QU'EST-CE QUE

LA CARTOGRAPHIE DU WEB ?

*Expéditions scientifiques dans l'univers
des données numériques et des réseaux*

FRANCK GHITALLA

Qu'est-ce que la cartographie du web ?

Expéditions scientifiques dans l'univers des données numériques et des réseaux

Franck Ghitalla

Dominique Boullier et Mathieu Jacomy (éd.)

DOI : 10.4000/books.oep.15358

Éditeur : OpenEdition Press

Lieu d'édition : Marseille

Année d'édition : 2021

Date de mise en ligne : 23 août 2021

Collection : Encyclopédie numérique

EAN électronique : 9791036562631

<https://books.openedition.org>

Édition imprimée

EAN (Édition imprimée) : 9791036562655

Nombre de pages : 288

Référence électronique

GHITALLA, Franck. *Qu'est-ce que la cartographie du web ? Expéditions scientifiques dans l'univers des données numériques et des réseaux*. Nouvelle édition [en ligne]. Marseille : OpenEdition Press, 2021 (généré le 03 février 2022). Disponible sur Internet : <<http://books.openedition.org/oep/15358>>. ISBN : 9791036562631. DOI : <https://doi.org/10.4000/books.oep.15358>.

Crédits de couverture

Conception graphique : Veronica Holguín

© OpenEdition Press, 2021

Creative Commons - Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International

- CC BY-NC-ND 4.0

RÉSUMÉS

Les *web studies* et les *network sciences* ont déjà une histoire, et, en France, cette histoire fut profondément marquée par la contribution de Franck Ghitalla.

Cet ouvrage posthume était en fait déjà prêt à être publié au moment du décès brutal de son auteur en décembre 2018. Ses chroniques du Web et ses carnets cartographiques, ici regroupés, ont, en effet, une vocation pédagogique certaine. Cet ouvrage est ainsi idéal pour une initiation, car il permettra à chacun de s'approprier les concepts et les méthodes que Franck Ghitalla a éminemment contribué à développer en France.

L'histoire de la construction de ces approches en réseau qu'il retrace dans la première partie permet de comprendre que la tâche requiert de multiples expertises : la rigueur des calculs doit être alliée à la fiabilité des sources et des données ainsi qu'à la sensibilité sémiotique de la visualisation pour rendre robuste une interprétation toujours ancrée dans les sciences sociales. Chaque chapitre s'appuie sur des exemples d'analyse de réseaux très précis, dans des domaines très divers, et les concepts sont progressivement distillés, ainsi que les tours de main, que Franck Ghitalla, brillant artisan des données numériques, avait à cœur de partager avec tous.

FRANCK GHITALLA

Franck Ghitalla a fait toute sa carrière à l'université de technologie de Compiègne, où il a été enseignant-chercheur en sciences de l'information et de la communication.

DOMINIQUE BOULLIER (ÉD.)

Professeur des universités en sociologie à l'institut d'études politiques de Paris (Sciences Po), chercheur au centre d'études européennes et de politiques comparée. Docteur en sociologie (EHESS), HDR en Infocom, diplômé de linguistique (Rennes 2). Spécialiste du numérique et en particulier des réseaux sociaux, des politiques des architectures numériques, des villes numériques, des méthodes numériques en sciences sociales (centrées sur la cartographie du Web puis sur les propagations). Il a dirigé plusieurs laboratoires de recherche (dont Costech à l'Université de Technologie de Compiègne, le Médialab de Sciences Po avec Bruno Latour, le Social Media Lab de l'EPFL, Lausanne) et créé des *user labs* (notamment Lutin, Cité des Sciences). Auteur de *Sociologie du numérique* (Armand Colin, Paris, 2016), de *Comment sortir de l'emprise des réseaux sociaux* (Le Passeur éditeur, Paris, 2020) et d'articles sur une théorie des propagations comme troisième génération de sciences sociales (blog : shs3g.hypotheses.org).

MATHIEU JACOMY (ÉD.)

Docteur en techno-anthropologie et concepteur d'instruments scientifiques pour les sciences sociales au TANT Lab de Copenhague. Ses recherches s'intéressent à la sociologie des sciences et des techniques, à l'analyse visuelle de réseaux, à la

cartographie des controverses et aux méthodes numériques. Il a contribué à créer le logiciel d'analyse de réseaux Gephi et l'outil de fouille du Web Hyphe. Il twitte à @jacomyma et publie un carnet de recherche anglophone sur reticular.hypotheses.org.

NOTE DE L'ÉDITEUR

Conformément à la licence CC BY-NC-ND, les images contenues dans cet ouvrage peuvent être réutilisées en dehors de toute utilisation commerciale, sans modification, avec mention des crédits.

QU'EST-CE QUE

LA CARTOGRAPHIE DU WEB ?

Cet ouvrage a été subventionné par
l'université de technologie de Compiègne

FRANCK GHITALLA

QU'EST-CE QUE

LA CARTOGRAPHIE DU WEB ?

*Expéditions scientifiques dans l'univers
des données numériques et des réseaux*

En couverture : *Close up of a spider web*
© your123 - stock.adobe.com
Conception graphique : Veronica Holguín

Direction scientifique : Dominique Boullier et Mathieu Jacomy
Suivi et coordination éditoriale : Julie Pynte et Cédric Gaultier
Correction : Solenne Louis
Mise en page grâce à Métopes. Méthodes et outils pour l'édition structurée.

Collection « Encyclopédie numérique », 2021

Cet ouvrage est en ligne en accès ouvert :
<http://books.openedition.org/oep/15358>

Texte : Licence Creative Commons Attribution – Pas d'utilisation commerciale
Pas de modification 4.0 International
<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.fr/>

ISBN papier : 979-10-365-626-55
ISBN électronique : 979-10-365-626-031

L'auteur

Franck Ghitalla a fait toute sa carrière à l'université de technologie de Compiègne, où il a été enseignant-chercheur en sciences de l'information et de la communication depuis 1999 et membre de l'unité de recherches Costech, à l'exception de trois années de recherche à l'INIST (2009-2011).

Docteur en linguistique de l'université Rennes 2, formé par Jean Gagnepain à la théorie de la médiation, Franck Ghitalla a fait sa thèse sur l'image, puis est devenu un pionnier des *web studies* en France et en Europe. Ses travaux sur la cartographie des données, les graphes du Web et sur la visualisation de communautés ou de thématiques en ligne ont abordé une grande quantité de domaines : information scientifique et technique, économie, politique, idéologies, cybersécurité.

Ses cours à l'UTC lui ont permis d'attirer un grand nombre d'élèves ingénieurs dont plusieurs ont fait carrière dans les *data sciences* et *web sciences*, en créant leur entreprise, comme Linkfluence ou l'Atelier Iceberg, ou des logiciels reconnus mondialement comme Gephi.

Son blog L'Atelier de Cartographie fut son principal moyen de diffusion et a fourni la matière première de cet ouvrage posthume. Franck Ghitalla est décédé brutalement le 19 décembre 2018, alors qu'il donnait son cours à l'UTC.

Les éditeurs scientifiques

Dominique Boullier est professeur des universités en sociologie à l'institut d'études politiques de Paris (Sciences Po), chercheur au centre d'études européennes et de politiques comparée. Docteur en sociologie (EHESS), HDR en Infocom, diplômé de linguistique (Rennes 2). Spécialiste du numérique et en particulier des réseaux sociaux, des politiques des architectures numériques, des villes numériques, des méthodes numériques en sciences sociales (centrées sur la cartographie du Web puis sur les propagations). Il a dirigé plusieurs laboratoires de recherche (dont Costech à l'université de technologie de Compiègne, le médialab de Sciences Po avec Bruno Latour, le Social Media Lab de l'EPFL, Lausanne) et créé des *user labs* (notamment Lutin, Cité des Sciences). Auteur de *Sociologie du numérique* (Paris, Armand Colin, 2016), de *Comment sortir de l'emprise des réseaux sociaux* (Paris, Le Passeur éditeur, 2020) et d'articles sur une théorie des propagations comme troisième génération de sciences sociales (blog : shs3g.hypotheses.org).

Mathieu Jacomy est docteur en techno-anthropologie et concepteur d'instruments scientifiques pour les sciences sociales au TANT Lab de Copenhague. Ses recherches s'intéressent à la sociologie des sciences et des techniques, à l'analyse visuelle de réseaux, à la cartographie des controverses et aux méthodes numériques. Il a contribué à créer le logiciel d'analyse de réseaux Gephi et l'outil de fouille du Web Hyphe. Il twitte à @jacomyma et publie un carnet de recherche anglophone sur reticular.hypotheses.org.

Préface

Mathieu Jacomy

Lorsque Franck nous quitta juste avant Noël 2018, fauché par un AVC alors qu'il donnait son légendaire cours de cartographie de la connaissance à l'université de technologie de Compiègne, le même cours que j'avais suivi comme étudiant ingénieur et où je l'avais rencontré, la France perdit son premier « cartographe d'information » et un passeur hors pair. Il avait le don d'éclairer. Enseigner et faire de la recherche étaient pour lui deux facettes d'une même exploration, et ses étudiants l'adoraient. Ce livre témoigne de la singularité de son entreprise intellectuelle : collective et empreinte de fascination, libertaire et subversive. Il en retrace l'aventure avec l'authenticité d'un artisan et la complicité d'un conteur.

Ce livre est né du blog de Franck, L'Atelier de Cartographie. Il y rendait compte du métier de « cartographe d'information » qu'il avait inventé à force d'explorations du Web et qu'il réinventait sans cesse. Il y décrivait sa propre pratique et celle des personnes qu'il avait inspirées et qui l'avaient inspiré. Retravaillant ses contributions, Franck écrivit et publia en ligne la version initiale de ce livre, chapitre par chapitre, jusqu'en 2017. Il avait pour projet de le finaliser pour en faire un ouvrage de papier, qui constitue de fait l'héritage qu'il nous laisse.

Ce livre raconte comment Franck est devenu « l'as des réseaux », pour reprendre le titre du portrait que le journal *Le Monde* lui consacra en 2012. Mais il était aussi l'as de son propre réseau d'étudiants et anciens étudiants, qui ont essaimé dans la recherche (un peu) et les *start-up* (beaucoup). Dans l'écriture comme dans la vie, Franck avait à cœur de montrer que ces échanges l'avaient nourri. Il dédaignait le surplomb du professeur, préférant apprendre *de* et *avec* les autres, enrichissant son talent de professeur du rôle d'apprenant qu'il endossait souvent pour absorber les techniques émergentes du Web et de la science des réseaux. Il était conscient que toute expédition scientifique naît d'un désir curieux, et les siennes avaient connu des succès qui avaient fait de lui un innovateur écouté. Non spécialiste, il trouvait des mots de passeur pour éclairer les nouvelles technologies. Ainsi qu'il l'écrit dans la seconde partie du livre, « les prouesses technologiques exploratoires sont les produits d'une forme d'ignorance ou de questionnement de départ, et non des garanties théoriques issues des mondes de l'érudition, de la répétition, des certitudes idéologiques ». Joignant le regard de l'expert à la perspective du profane, Franck raconte les liens intellectuels, pratiques et humains qu'il a tissés entre données et connaissance, écriture et technologie numérique. Il questionne le Web, les réseaux sociaux et d'autres réseaux encore, expérimentant de nouveaux moyens de les visualiser et de les analyser.

Franck a écrit son ouvrage en deux livres distincts, avant de se raviser et les réunir en un seul. La seconde partie, « Études cartographiques », donne une série d'exemples du travail de cartographe d'information, qu'il présente comme « une des formes de l'artisanat de haute technologie dédié aux data ». Dans ces cas d'études, Franck le cartographe aguerri expose et interroge sa pratique sous différentes formes. Par contraste, la première partie, « Chroniques du Web », raconte la genèse et les hésitations de son entreprise. Franck l'explorateur nous montre d'où viennent les « instruments cartographiques » qui lui seront nécessaires. On y comprend que le Web fut l'archétype d'autres espaces de connaissance à cartographier.

C'est à cette époque que nous avons le plus collaboré, construisant des outils de collecte de données web (le Navicrawler) et d'analyse de réseaux (Gephi) dont il était un utilisateur averse et perspicace. Dans son texte, Franck porte un regard bienveillant, mais critique sur cet appareillage. Il témoigne notamment de la fascination qu'exerce la cartographie de l'information, qui ne lui était pas étrangère et dont il savait tirer parti pour convaincre. Il avait absorbé sans filtre les promesses d'Albert-László Barabási, pionnier états-unien de la science des réseaux, fameux vulgarisateur, controversé pour avoir prématurément proclamé l'universalité des lois de la complexité¹. Franck témoigne dans ce livre de l'importance de Barabási dans la construction de sa pratique cartographique. Il avait fait sienne la fascination de l'universalité, qu'il traduisait par la promesse d'une perspective en plongée sur la connaissance, d'un point de vue aérien. Mais il n'ignorait pas « l'effet de réalité » des cartes réseau. Sa pratique cartographique était ambiguë, tantôt arme de persuasion et tantôt mesure défensive contre la trahison des images. À la cohérence théorique, il préférait la multiplicité d'expérimentations incarnées : collectives, empiriques, et matérielles. « C'est dans ces moments-là que l'on mesure le pouvoir heuristique de la cartographie d'informations, lorsqu'elle disparaît à titre de filtre ou d'instrument. Notre alignement de processus de traitement des data ne fonction[ne] donc qu'en devenant transparent. Invisible. » Franck s'était réapproprié la vision des pionniers de la science des réseaux, et de leur expérience il avait retenu la leçon de l'hostilité du monde académique envers ceux qui transgressent ses règles. Il avait donc pris le parti d'emprunter des chemins de traverse, refusant de se soumettre à la publication dans des revues à comité de lecture, échappant à des critères d'évaluation qu'il jugeait arbitraires, leur préférant un rapport direct au public qu'il s'était constitué.

Franck ne s'embarrassait pas d'entraves à sa créativité. Délaisant l'habit du savant pour le tablier de l'innovateur, il avait entrepris de « subvertir » la tradition académique. Il l'explique dans son livre : « Personnellement, je me suis toujours un peu

1. M. Jacomy, "Epistemic clashes in network science: Mapping the tensions between idiographic and nomothetic subcultures", *Big Data & Society*, 2020, doi:10.1177/2053951720949577.

ennuyé dans cette tradition spéculative et érudite, celle pourtant d'où je viens à titre de chercheur en sciences humaines et sociales. Je préfère aujourd'hui partager l'enthousiasme (et la naïveté sûrement) de ceux qui cherchent volontairement à briser les équilibres dont nous avons hérité entre hypothèse théorique et dispositif expérimental, ce qu'avait justement mis longtemps à polir la tradition ». Mais paradoxalement, son énergie libertaire inspira de nombreux chercheurs et contribua à renouveler les pratiques des humanités francophones. Il faut dire que sous la surface quantitativiste de l'analyse de réseaux, la perspective de Franck devait moins à la physique statistique d'un Barabási qu'à l'archéologie de la connaissance d'un Michel Foucault. Son point de départ avait simplement été de contempler et de cataloguer le Web selon le mode de connaissance de l'âge classique, tel que raconté par Foucault dans *Les mots et les choses*. C'est l'universalité de ce projet, c'est-à-dire le rapport descriptif à son objet (le réseau), qui lui permit d'inventer une collaboration pratique avec ses étudiants ingénieurs. Franck sut défricher une voie transdisciplinaire intelligente qui irrigua notamment les humanités numériques. Son livre en donne les clés.

Puisque je joue dans cette histoire le rôle d'un faiseur d'instruments, je dois souligner l'importance pour quelqu'un comme moi de collaborer avec quelqu'un comme Franck. On pourrait penser qu'un logiciel populaire comme Gephi a été développé pour un large public, mais ce n'est pas le cas. Pour des raisons pragmatiques, les retours d'un petit nombre d'utilisateurs engagés sont préférables à de nombreux retours d'usage superficiels. Il s'ensuit une chose qui ne va pas de soi : nos instruments n'ont pas eu pour but d'être utiles à des débutants, mais à des personnes capables d'inventer et d'interroger leurs pratiques. Même si les algorithmes et techniques embarqués dans cet appareillage restaient partiellement hermétiques à Franck, il en comprenait les enjeux mieux que la plupart des informaticiens et pouvait en formuler une critique heuristique, enracinée dans une expérience épaisse. Il passait tout instrument à l'épreuve opérationnelle, sondant son utilité pratique et sa capacité à produire du sens. La plupart des instruments qu'il contribua à façonner se disséminèrent d'eux-mêmes parce qu'ils étaient fructueux, c'est-à-dire générateurs de connaissance, ce qui les rendait parfois subversifs.

On entend dire que cartographier l'information s'apparente à lire dans le marc de café. C'est vrai, et le concevoir comme une critique serait méconnaître l'importance de la divination dans la genèse de l'écriture². Mais les mots nous trompent, car, s'il s'agit de lire les signes, il ne s'agit en rien de deviner quelque chose. Il faut au contraire comparer et questionner les signes de la carte pour entrevoir la configuration matérielle

2. E. Souchier et P. Judith, « Voir le web et deviner le monde. La "cartographie" au risque de l'histoire de l'écriture », dans B. Galinon-Mélénec et S. Zlitni (dir.), *Traces numériques. De la production à l'interprétation*, Paris, CNRS éditions, 2013, pp. 213-234.

des données. La cartographie d'information est une écriture pratiquée comme une lecture où la manipulation des signes est en partie déléguée à l'instrument, à l'algorithme de « spatialisation » pour reprendre le nom que Franck lui donna. Elle fut jugée subversive par des positivistes. Pourtant cette technique de connaissance et ses cousines, les cartes auto-adaptatives et la réduction de la dimensionnalité, se popularisèrent. On les utilise dans un contexte exploratoire où la question de leur objectivité n'est pas problématique, en complément d'autres approches. Dans la tradition de l'analyse exploratoire de données initiée par John Tukey, elles ont pour rôle de faire naître des hypothèses que l'on validera par d'autres moyens. Comme la divination, les techniques de connaissance développées par Franck brillent par leur capacité à faire naître du sens. Il concevait son travail comme celui qui consiste à apporter la lumière dans la profondeur aveugle des données, et son livre accompagnera ceux qui partagent son entreprise.

Prologue

Ce livre n'est pas un ouvrage scientifique, du moins au sens classique. Il s'agit plutôt du récit d'une quête curieuse animée par une question qui m'obsède : le Web, ce vaste système distribué de documents et de contenus, peut-il avoir une *forme* ? Peut-on en décrire l'organisation, les propriétés, la *géographie* ? Le Web, la couche la plus abstraite de l'Internet, peut-il être décrit comme un *espace* ou un *objet*, un *territoire* ou un *processus* ?

Dès mon arrivée à l'université de technologie de Compiègne en 1997 comme chercheur en sciences de l'information, j'ai intégré le département Technologie et sciences de l'homme, ce genre de département qu'affectionnent les écoles d'ingénieurs, censés apporter aux étudiants ce « supplément d'âme » nécessaire à la compréhension des enjeux politiques, économiques ou sociaux des technologies. J'ai donc commencé par enseigner la « linguistique et la philosophie du langage », puis, dans un second temps, les « usages des technologies numériques de l'information ». Tout aurait dû me conduire à développer des recherches depuis cette « périphérie de la technologie » où l'on examine ses usages sociaux, les opérations cognitives qu'elle supporte, les interfaces par lesquelles elle nous apparaît, l'histoire et l'épistémologie qui en décrit les périodes ou l'anthropologie qui en étudie les différents rôles. Un réseau ouvert de documents numériques comme le Web aurait donc pu constituer pour moi un excellent terrain d'étude du point de vue des sciences humaines et sociales : un terrain d'observation encore vierge à l'époque que j'aurais décrit avec ce regard critique nourri d'érudition qui fait la spécificité des SHS. Oui, mais voilà : au début des années 2000, lorsque je commençais à enseigner en école d'ingénieurs, le Web entamait sa formidable expansion et posait des questions massives auxquelles il allait falloir se confronter et auxquelles n'étaient pas préparées les sciences humaines et sociales, ni même peut-être nos universités : ses masses, son caractère distribué et son évolution temporelle, ses technologies, le modèle d'innovation qui sous-tend une telle dynamique de développement... et les raisons des premiers succès des géants californiens de l'information qui dominent encore aujourd'hui le monde des moteurs de recherche et des services. Faute d'ingénierie spécifique et de programmes technologiques ambitieux en France comme en Europe, les sciences humaines n'ont pas su inventer les instruments originaux et l'assise expérimentale qui aurait permis de nourrir de leurs spécificités toutes les innovations qui sont sorties de ce vaste univers de *data* et de services. L'idée de *produire* la technologie plutôt que de la *décrire* s'imposa donc clairement à moi et, si le Web posait autant de questions, autant tenter d'y répondre *de l'intérieur*.

Cet univers m'intriguait d'autant plus que tout le monde reconnaissait l'importance de ce réseau, mais s'évertuait à le représenter sous la forme d'un nuage énigmatique. Au printemps 1999, j'entrepris donc de modifier la forme de mon enseignement et ses objectifs pour me concentrer sur un défi dans lequel les élèves ingénieurs me suivirent naturellement : « dessiner le Web » ! Tout fut bon pour arriver à un schéma ou un plan intelligible plutôt qu'un nuage indistinct : défricher la littérature scientifique sur le sujet, suivre de près les développements technologiques dans l'univers des premiers moteurs de recherche et, surtout, construire nous-mêmes ces petits programmes qui permettent de naviguer sur le Web et de récolter nos propres jeux de données, les *web crawlers*. Petit à petit, je me mis à valider les étudiants autant par le « faire » que par le « dire » ou l'écrit et, à ce jeu-là, tout me convenait qui puisse ouvrir l'imagination à cette sorte de *cybergeography* du Web : expériences, analyse de données, algorithmes, mais aussi plans, schémas et même maquettes ! En découvrant les premières réalisations des étudiants, j'acquis la conviction que l'effort exigé pour représenter graphiquement le Web constituait une étape précieuse pour développer ensuite les technologies et les services qui seraient adaptés à sa morphologie, à son architecture, à sa géographie documentaire. Une porte ouverte sur l'innovation.

Mon enseignement ressemble depuis lors à un atelier de créativité d'où sont sortis des projets qui ont rythmé ma vie, et nourri aussi quelques convictions. La première est qu'il est souvent bien difficile de trouver une place à l'innovation dans nos universités et nos écoles, toujours coincées entre « enseignement » d'un côté et « recherche » de l'autre. En modifiant mon enseignement, je transformai donc aussi mes pratiques de chercheur pour construire cet espace nécessaire à nos nouvelles activités où nous avons emprunté de nombreux chemins et où le hasard eut sa place. La seconde est qu'il ne faut pas limiter l'innovation aux cadres de « l'innovation industrielle », comme c'est trop souvent le cas en école d'ingénieurs. Comme le démontrent aujourd'hui les technologies web dans leur transformation du monde, l'innovation peut être tour à tour, et sans hiérarchie, « industrielle », mais aussi *scientifique, sociétale, culturelle* ou *politique*. Il faut donc admettre que l'innovation technologique ne se développe efficacement que dans un cadre « d'enseignement » ou de « formation » profondément réinventé. De là, une troisième conviction : former des ingénieurs (et peut-être au-delà, toute la jeunesse), c'est admettre avec eux une forme *d'ignorance partagée* sur quelques-unes des problématiques fondamentales qui marquent notre temps, comme les problèmes climatiques et environnementaux, la biodiversité en danger, les méfaits de l'économie spéculative ou... les zones d'ombre qui masquent encore quelques-unes des propriétés étonnantes du Web comme sa dynamique évolutive. Soyons francs : face à de tels défis théoriques ou sociétaux, et à relever si rapidement si nous ne voulons pas en payer le prix fort, peut-on faire autre

chose que de partager notre ignorance avec une jeunesse dont l'appétence pour le monde de demain est sans limite ?

La forme de cet ouvrage peut étonner : une série de chapitres rédigés de façon libre et personnelle où le *je* comme le *nous* alternent souvent indistinctement, pour marquer la dimension collective d'une série d'aventures expérimentales. Pour parler vrai, ni le « je » ni le « nous », et encore moins le genre de recherche technologique dont il est question dans cet ouvrage, n'ont vraiment de place dans la recherche scientifique française, cloisonnée par ses disciplines, académique dans ses procédures, figée dans son organisation, conservatrice dans ses formes d'expression. C'est pourquoi on ne trouvera guère dans les revues officielles nos contributions, ces fameux « articles » dont le nombre et la supposée qualité sont censés m'évaluer à titre individuel (tout à fait réglementairement puisque je suis fonctionnaire) et constituer des indices fiables de l'*excellence scientifique*. Outre le caractère absurde de l'expression en science, les textes officiels nous enjoignent de « rayonner » à travers les canaux restreints de la publication académique, sans tenir compte, à aucun moment, de toutes ces externalités par lesquelles nous diffusons les démarches, nous renouvelons les méthodes, nous inventons les instruments de demain, nous créons de l'activité et des emplois. *Rayonné-je* ? La question m'apparaît aussi ridicule que sa prononciation. Et puis rayonner dans quel type d'espace ? Je ne sais toujours pas aujourd'hui à quel type exact de recherche nous nous sommes livrés pendant des années, et pas plus aujourd'hui. Notre bricolage inventif, fondé sur une ignorance partagée à l'égard d'un réseau qui semble encore échapper à notre compréhension, se situe pourtant quelque part entre « recherche » et « enseignement » dans nos universités, entre sciences humaines et sciences informatiques, mais sans que je n'aie jamais pu rattacher notre projet aux unes ou aux autres. Un *quelque part* qui n'existe, à vrai dire, ni dans les programmes de recherche européens qui petit à petit nous sont imposés (le fameux H2020, Horizon 2020 de la recherche européenne), ni dans les programmes régionaux et encore moins dans les programmes redondants de l'Agence nationale de la recherche (ANR) dont les financements sont systématiquement accordés aux mêmes équipes depuis toujours.

Le Web fut donc salvateur à bien des égards. Dès le départ de nos aventures, dans les années 2002-2003, il a constitué notre objet d'étude favori, mais aussi un excellent vecteur de diffusion de nos trouvailles techniques et méthodologiques. Je ne compte plus aujourd'hui le nombre de fois où des visualisations de graphes issues

de Gephi (notre application favorite) ont été publiées dans *Science*, *Nature*, *National Geographic*, *Forbes* ou le *Washington Post* à propos de Social network analysis (SNA), mais aussi de génétique, de biologie, de sciences de l'environnement, d'économie ou de *data sciences*. Je n'ai jamais non plus identifié précisément la façon dont nous avons pénétré la recherche américaine à Stanford, au Massachusetts Institut of Technology (MIT) ou chez Facebook ou Google, mais le réseau a joué un rôle fondamental ici aussi. Je ne connais seulement, ou simplement me rappelle, que les destins personnels de ceux qui ont contribué à dessiner ce qui ressemble à une sorte d'écosystème dont le Web est l'ossature : les fondateurs de Linkfluence, le consortium Gephi, ceux partis contribuer à la R&D de LinkedIn ou d'autres entreprises en Californie, la création de Linkurious, la librairie SigmaJS, certains outils du médialab de Sciences Po comme le Navicrawler, la naissance récente de l'Atelier Iceberg. Nous avons donc étudié, mais aussi nourri de nos instruments l'écosystème du Web. Ce livre en est issu, presque directement puisqu'une partie des informations contenues dans cet ouvrage sont extraites de mon blog, L'Atelier de Cartographie¹. Ouvert en 2011, le blog a accueilli depuis lors plus de 42 000 visiteurs uniques et plus de 140 000 visites. Je suis loin, évidemment, du million de téléchargements de Gephi rien qu'en 2014 ! Accumulés chronologiquement, mes posts ont déjà tout d'un récit qu'il s'agissait de mettre en forme à travers cet ouvrage.

L'Atelier de Cartographie archive aussi les quelques documents que j'ai produits pour décrire les pistes possibles de développement d'instruments inédits et de méthodes nouvelles en ingénierie de l'information ; ils étaient essentiellement programmatiques et tournés vers les applications possibles de nos travaux. Ce type de document et de « posture » comme on dit n'a pas sa place dans la littérature scientifique officielle, y compris en école d'ingénieurs, mais peut s'avérer précieux comme source d'inspiration. C'est pourquoi l'idée d'écrire cet ouvrage m'est venue dès 2008 lors de la publication de l'un de mes rares articles dans une revue scientifique, précisément *Communication et Langues* n° 158 de décembre de la même année. Mon article était intitulé « L'Atelier de Cartographie » et consacré à la Toile européenne, une étude cartographique des sites web consacrés à l'Europe réalisée par une jeune *start-up* de l'époque, Linkfluence, qui travaillait en partenariat avec le Centre d'information sur l'Europe. Le responsable de la revue publie dans le même numéro un article consacré à l'écriture scientifique, à sa nature et à ses règles, pour signifier combien le mien n'entraîne pas dans les cadres convenus de l'exercice. Comme l'écrit le directeur de publication dans ce numéro, mon article a fait « l'objet d'échanges animés dans le comité de lecture ». Et il poursuit : « Deux types de questions ont avant tout cristallisé l'attention au cours de ces échanges. Le premier a porté sur la pratique d'écriture des textes scientifiques ; c'est une interrogation

1. <http://ateliercartographie.wordpress.com>.

récurrente dans l'équipe de *Communication et Langages* et plus généralement dans les publications en sciences humaines, l'écriture de certains articles étant parfois jugée trop subjective, voire trop littéraire, par certains lecteurs. Le second type de questions a quant à lui porté sur la posture énonciative des textes scientifiques. À l'origine de ces remarques, l'article sur l'Atelier cartographique consacré aux pratiques et aux enjeux de la "cartographie" thématique de documents web ». Il fallait donc trouver d'autres moyens de diffusion de ce que je pense être les sciences *de* l'ingénieur, et non pas des sciences *pour* l'ingénieur. Et, quitte à déroger aux règles du conservatisme académique, autant le faire sous forme d'un récit.

* * * * *

Sur ce point, je/nous ne faisons que poursuivre le chemin ouvert par des ouvrages légendaires qui ont façonné ce domaine inédit de la *science des réseaux (network sciences)* dans lequel notre projet de carte du Web puise son inspiration et une bonne partie de ses méthodes. Plongés dans notre bricolage intellectuel et technique, ils nous ont éclairés à bien des égards en remettant en perspective les nombreuses contributions scientifiques produites à partir de la fin des années 1990 en matière de *web sciences*, puis de *network sciences*. Nous avons ainsi pu coller, littéralement, à la recherche américaine où se sont hybridées la recherche scientifique et l'ingénierie des *data* donnant naissance à ce que l'on appelle désormais les *data sciences* avec son cortège de déclinaisons comme l'*open data*, les *web sciences*, le *social data mining* ou le *big data*. Ils se sont succédé comme des balises, orientant souvent nos objectifs : *The Laws of the Web* de B.A. Huberman en 2001, *Nexus* de M. Buchanan en 2002, *Linked* de A.-L. Barabási et *Six Degrees* de D. Watts en 2003, *Sync.* de S. Strogatz en 2004, *The Structure and Dynamics of Networks*, l'ouvrage collectif publié en 2006, *Networks* de M.E.J. Newman et *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* de J. Kleinberg et D. Easley en 2010. Nous n'avons jamais eu le temps, ni peut-être l'envergure intellectuelle, pour contribuer nous aussi au plan théorique à ces nouvelles sciences fondées sur l'ingénierie des données autant que sur les modèles théoriques. Sur la trentaine d'étudiants ou de jeunes ingénieurs qui ont travaillé avec moi, seulement trois se sont inscrits en doctorat, et deux seulement ont aujourd'hui soutenu leurs thèses. À titre de chercheur, je le regrette, mais j'ai partagé avec eux cette envie d'aller plus vite pour épouser l'innovation technologique accélérée des Californiens, plus loin pour réinventer les méthodes et les instruments associés aux masses de *web data*. Une partie des jeunes ingénieurs qui m'ont accompagné partagent encore aujourd'hui cette curieuse quête de la forme du Web à travers leurs activités professionnelles, chacun à sa façon.

Nous avons donc organisé, dès le début, un modeste atelier artisanal, une sorte de laboratoire collectif hors les murs où nous avons accumulé les projets *ad hoc* dont chacun a été l'occasion d'enrichir notre boîte à outils méthodologiques et nos instruments. Notre véritable trésor de guerre. Nous n'avons jamais eu d'organisation formelle, hormis l'association WebAtlas qui a vécu quelque temps. Ainsi qu'un document d'une page rédigé en décembre 2000 intitulé RTGI pour Réseaux, territoires et géographie de l'information. En revanche, nous avons toujours revendiqué notre rôle dans ce que j'appelle désormais *l'artisanat de haute technologie* : dans l'univers de l'innovation autour des *web data*, ce sont toutes ces chaînes de traitement de l'information qui, depuis l'extraction des données jusqu'aux interfaces finales, ouvrent sur l'idée de nouveaux services, de nouvelles activités et, je l'espère, de nouveaux types d'emplois ou d'activités. Au cours des années, je me suis spécialisé dans le développement et l'usage intense des instruments cartographiques. N'étant pas suffisamment compétent en mathématiques et en sciences informatiques (je ne suis pas ingénieur !), certains domaines des chaînes de traitement de l'information m'étaient fermés comme le codage d'algorithmes ou la conception d'architectures système. Je me suis donc tourné, petit à petit, vers les instruments de cartographie de l'information ou, pour être plus précis, de *cartographies de graphes relationnels*. Les graphes sont des instruments d'analyse des phénomènes complexes sous l'angle de leurs aspects systèmes (comme peut l'être l'analyse multidimensionnelle). Les aspects mathématiques liés aux matrices de graphes peuvent être d'une grande abstraction (la *théorie des graphes*), mais on peut leur associer aussi des instruments de spatialisation graphique, des sortes de *topographies* ou de *cartographies* qui accompagnent les propriétés statistiques alors incarnées visuellement. Le Web, à titre de système documentaire ouvert et largement distribué, se prête à merveille à une analyse à base de graphe : les pages web peuvent être désignées comme des *nœuds* reliés entre eux par des *liens* de différentes natures, dont le lien hypertexte qui constitue la colonne vertébrale du réseau. L'infrastructure logistique des moteurs de recherche web fournit, par ailleurs, les bases ou les archives nécessaires aux calculs des propriétés statistiques d'immenses masses d'information. C'est ce passage à l'échelle permis par le Web et ses technologies qui, à mon sens, a profondément renouvelé la théorie des graphes et ses instruments mathématiques et techniques.

Au début des années 2000, je me suis donc retrouvé dans une situation inédite avec d'un côté, les masses de données web désormais accessibles puisque le réseau est ouvert à l'exploration et, de l'autre, les premières applications de visualisation de graphes dont Gephi a désormais marqué l'histoire. Je suis donc devenu une sorte de « photographe du Web » pendant plusieurs années, avant d'aborder plus récemment d'autres univers documentaires. Pour moi, ce furent autant d'expéditions dans l'univers des données en réseau, sur le Web (*Chroniques du Web*, première partie de

l'ouvrage) ou ailleurs (*Études cartographiques*, seconde partie). Et j'occupe encore aujourd'hui cette place indéfinie de cartographe de l'information, une place qui commence là où finissent les données et s'achève là où commence l'interprétation experte du domaine ainsi cartographié. Un espace où se sont déroulées la plupart de nos explorations de l'univers des *web data*, et dont j'ai rapporté quelques clichés regroupés dans cet ouvrage.

Chroniques du Web

Chapitre I. Voir le Web

« The discovery of scale-free networks induced a paradigm shift: They taught us that the many complex webs surrounding us are far from random, but are characterized by the same robust and universal architecture. I am repeatedly asked a few basic questions when I lecture about networks: Why did it take this long?

Why did we have to wait until 1999 to discover the impact of hubs and power laws on the behavior of complex networks? The answer is simple: *We lacked a map* », Albert-László Barabási, *Linked*, 2003.

Lorsque je finis de lire en 2001 “Graph Structure in the Web”, l’article retentissant d’Andrei Broder et de ses collègues, la question s’est imposée avec évidence : le Web pouvait-il avoir une *forme* ? Peut-on en produire une *image* ou un *modèle* et, si oui, selon quels procédés, selon quelles méthodes, avec quels instruments ? Je me souvins alors qu’il existait déjà des travaux connus sur la cartographie de l’Internet comme structure physique reliant par câbles différents points du globe, ou alors comme réseau distribué de *routers* par où transitent les masses d’informations¹. Mais de là à imaginer que l’on puisse cartographier le Web dans ce qu’il a de plus abstrait à titre d’*espace de publication de documents*, d’images et de sons, il y avait manifestement un pas qui venait d’être franchi. Une géographie documentaire...

L’article du groupe de chercheurs visait à étudier quelques-unes des propriétés statistiques du Web sous la forme d’un graphe contenant plus de 200 millions d’éléments (les pages web étant considérées comme des *nodes*) reliés entre eux par plus d’un milliard et demi de liens (les liens hypertextes entre les pages ou *links*). Rien que cela ! Pour ceux qui ne maîtrisent pas les mathématiques statistiques, il est difficile d’apprécier les méthodes développées dans l’article.

1. Center for Applied Internet Data Analysis, <http://www.caida.org/home/>.

Mais on aperçoit rapidement que les auteurs cherchent à traduire les masses d'informations accessibles sur le Web en un ensemble de propriétés physiques, comme s'il s'agissait d'un espace mesurable, gouverné par des propriétés matérielles comme le *diamètre*, la *densité*, les *composants* et leur *hiérarchie*. Autrement dit, toute une géographie d'un nouveau genre dans laquelle les liens hypertextes distribuent les documents dans des *clusters*, entre lesquels s'exercent des *forces d'attraction*, incarnés par des *flux* de liens entrants et sortants dessinant ainsi des *distances*, des *zones*, un *centre*, une *périphérie* et même des *excroissances* !

Image 1.2

Le modèle du Web en nœud papillon, d'après : A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins et J. Wiener, "Graph Structure in the Web", *Computer Networks*, 33 (1-6), 2000, pp. 309-320.

L'article est passé à la postérité pour le schéma du Web « en nœud papillon » qui y figure, et l'on ne compte plus aujourd'hui les ouvrages qui le reprennent tel quel comme une sorte de « plan » général du réseau. L'idée a de quoi fasciner : plus de 200 millions de documents y sont placés dans des zones identifiables, chacune d'elles gouvernée par la façon dont les liens s'y distribuent. Autour d'un ensemble central, très dense de connexions internes, s'étirent à gauche et à droite deux « oreilles » qui comportent à peu près le même nombre de pages web, à la différence majeure que l'ensemble de gauche génère des liens qui pointent vers l'ensemble central et celui de droite reçoit des liens qui en proviennent. Le Web semble massivement gouverné par une sorte de magnétisme massif, accompagné de quelques particularités remarquables comme ces excroissances (*tendrils*) sur les pôles de gauche et de droite, des *tubes* qui mènent directement de l'un à l'autre (autrement dit, on peut passer de l'un à l'autre via des chemins hypertextes de documents à documents) et quelques *disconnected components* qui posent question dès qu'on aperçoit que des « bouts du Web » semblent fonctionner de façon autonome (est-ce l'effet d'un corpus

parcellaire puisque ce ne sont que des données partielles, un échantillon ? Ou bien y a-t-il réellement des « territoires autonomes », cachés au reste du Web, parce que non connectés ? Et, si oui, que recèlent-ils en termes de contenus particuliers ? Peut-être des entrées sur ce que l'on appelle aujourd'hui le *dark web*...).

Ce plan schématique avait tout d'un genre inédit de boussole, une synthèse puissante qui laissait entrevoir des propriétés essentielles de ce réseau qui engouffrait autant de données et qui commençait, en ces années 2000-2003, à tisser sa toile jusque dans nos maisons et nos objets quotidiens. Rétrospectivement, le *nœud papillon* semblait représenter une promesse, un projet, un défi : je compris en l'étudiant qu'il ne s'agissait pas seulement d'un exercice de synthèse où se trouvait articulé graphiquement l'essentiel des propriétés statistiques de l'étude, comme on en trouve souvent dans les publications scientifiques. Cette figure spatialisante du Web annonçait le moment où deviennent nécessaires des modèles opérationnels d'un réseau dont on commence à apercevoir les propriétés, ces « *patterns* robustes » ou ces « lois » qui en règlent l'architecture comme l'évolution. Une fois ses propriétés essentielles identifiées, une foule de technologies pouvait être développée et mobilisée pour l'exploiter. Une source d'innovation inépuisable pour ceux qui ont commencé les premiers à dresser cette carte du réseau, les pionniers de la Silicon Valley. La publication de ce modèle graphique et logique du Web était le premier pas de l'exploration et de l'arpentage méthodique de ces territoires numériques peuplés d'informations organisées en réseaux. Et participer à cette expédition eut tout pour moi d'une évidence, d'autant plus que j'étais entouré de jeunes ingénieurs qui se sont plongés dans l'observation de ce vaste système d'informations distribuées qu'est le Web et dont très peu de propriétés étaient connues avant les années 2000. Mais observer le Web, le *voir*, n'avait rien d'une évidence : les instruments dont nous avons besoin devaient être développés et, pour partie, inventés. Le genre de *photographie* que nous espérions pouvoir produire supposait la mise en œuvre de techniques et de méthodes conçues pour un univers ouvert et distribué, et d'une croissance exponentielle. C'est ce genre de défi qu'aiment à relever parfois de jeunes ingénieurs, être en limite de la science pour ouvrir de nouvelles voies en termes de technologies dédiées aux *data*, de méthodes de calcul ou de réflexions théoriques sur la *topologie* des systèmes d'information. De nouvelles voies, certes, mais aussi une série de décalages, voire de ruptures avec les cadres habituels à partir desquels étaient pensés jusqu'alors les systèmes d'information ou les systèmes de connaissance.

L'idée de *voir le Web* fit vite office d'objectif commun de recherche technologique pour les jeunes ingénieurs qui travaillèrent au lancement de notre programme d'expérimentations. Cet enthousiasme s'expliquait par le fait qu'il devenait enfin envisageable d'explorer la structure réelle du Web comme une architecture documentaire à part entière et dont nos usages quotidiens ne révèlent qu'une vue très parcellaire. Usuellement, l'accès à cet espace dépend le plus souvent de ces deux instruments que sont les navigateurs, puis les moteurs de recherche qui nous renvoient leurs résultats sous forme de listes ou d'éléments ordonnés selon différents critères. Nos accès au Web comme nos activités passent essentiellement par ces deux *filtres*, pratiquement transparents et naturalisés pour les internautes. Il nous fallait donc *passer de l'autre côté*, oublier le Web comme *instrument* ou *vecteur* de production d'informations ou de connaissances pour en faire un *objet d'investigation*, assuré comme phénomène observable, attesté dans ses frontières matérielles. Il fallait donc réunir des instruments d'observation ou d'inspection, les imaginer et les développer si nécessaire, pour apercevoir *nous aussi* ce vaste système distribué de documents et des flux qu'y dessinent les liens hypertextes.

Certes, il s'agit encore d'*instrumentation*, donc d'un autre type de filtres tout aussi artificiels que le navigateur et le moteur de recherche. Mais cet appareillage de traitement des données serait tourné cette fois sur l'arrière-cour du système, de son ossature informationnelle, des lois qui gouvernent son architecture. Et quelle architecture ! Dès le départ de nos premières expérimentations, elle se révéla fascinante, mais aussi déstabilisante pour des héritiers de la culture du XX^e siècle et de ses instruments de prédilection comme peut l'être une bibliothèque (les livres ordonnés par dates, thèmes ou auteurs) ou une arborescence (les éléments distribués à des niveaux stricts de hiérarchie). La visualisation des connexions de document à document (ou de page web à page web) laissa rapidement apparaître un nouveau type d'architecture documentaire, une géographie foisonnante dont il allait falloir extraire des lois fondées sur le nouveau principe de *proximité* qu'induisait le lien hypertexte, et sa distribution à grande échelle. À ce titre, le langage HTML à partir duquel est construit le document web (ou la « page ») offre l'avantage remarquable d'agrégier des contenus entre eux, mais aussi d'insérer dans les balises des *ancres* et des *cibles* qui sont écrites, donc accessibles depuis la chaîne de caractères. On pouvait donc relever automatiquement la liste des connexions de page à page, sans aucune idée des *patterns* qui allaient apparaître aussitôt passés certains seuils quantitatifs. Les *web crawlers* ont ainsi joué un rôle majeur dans nos premières expéditions sur le Web et ils concourent encore aujourd'hui à analyser les documents accessibles sur le réseau pour les moteurs de recherche, et à repérer la publication de nouvelles sources d'information... ou leur disparition.

Il est d'ailleurs important de souligner combien les moteurs de recherche ont contribué à traduire les données web en instruments presque « naturels » pour les usagers. Ils affichent leurs résultats sous forme de *listes* et pas de *cartes*, ils intègrent des principes de classification avec leurs catégories de recherche et restent basés sur l'extraction des « contenus » textuels ; les algorithmes parmi les plus connus, comme le *pagerank*, s'inspirent des techniques d'analyse des citations en bibliométrie qui les précèdent historiquement. Ces aspects nourrissent finalement l'idée que le Web est un produit de l'univers de la bibliothèque, sa *filiation* dans le monde numérique. Oui, mais le modèle d'organisation que décrivait l'article d'Andrei Broder et ses collègues où figurait le schéma en *nœud papillon* ne correspondait à rien de connu et, si l'on pouvait tirer de ce schéma un instrument d'aide à la navigation sur le Web, il n'aurait rien de commun à une liste ou à une arborescence.

Le projet d'analyser la distribution de la connectivité (parallèlement à l'indexation des contenus) constituait une voie ouverte depuis longtemps comme je l'appris petit à petit. Le *nœud papillon* ne venait en réalité que couronner une série de travaux expérimentaux entamés plusieurs années auparavant. Le principe de la mesure des espaces en différents types de réseaux à partir de la distribution des *liens* ou des *connexions* a représenté une voie grandissante du domaine du *data mining* puis du *web mining*, qui donnera d'ailleurs le jour aux *web sciences*. Dès la naissance des premiers systèmes hypertextes (avant le Web), on a appliqué ainsi aux données réseau une série de techniques issues de la tradition de la théorie des graphes, cette branche des mathématiques appliquées qui cultivait déjà un intérêt certain pour les méthodes de projection graphique. Mais l'arrivée de données massives sur la distribution des liens hypertextes entre les pages avait permis de franchir un cap décisif : on pouvait enfin apercevoir des principes d'organisation générale ou, plutôt, des *régularités statistiques fortes*² comme certaines formes de hiérarchisation entre les documents à moyenne échelle (quelques milliers d'éléments) ou encore des différences de densité hypertexte repérables dans les corpus.

Ces premières avancées illustraient déjà la nécessité d'être équipé techniquement, entouré de compétences avancées et convaincu que l'exploration du Web prendrait la forme d'un lent arpentage de sa structure, comme s'il s'agissait de reconstruire le guide de la bibliothèque en parcourant soi-même, et petit à petit, tous ses rayonnages et leurs livres. L'image est à peine exagérée et résumerait presque l'essentiel des technologies d'exploration du Web développées depuis les années 1990 : les *crawlers* qui « sautent » de lien hypertexte en lien hypertexte pour dénombrer les pages en suivant les fils de la « toile », les bases où les contenus comme les liens

2. B.A. Huberman, *The Laws of the Web: Patterns in the Ecology of Information*, Cambridge (MA), MIT Press, 2001.

sont archivés, les algorithmes de traitement de contenu et les calculs associés qui permettent de faire émerger des principes de classification, de regroupement en *clusters*, de hiérarchisation des pages selon des règles de *ranking* qui vont être retournées à l'internaute lors de ses requêtes. Nombre de ces technologies ont contribué et contribuent encore aujourd'hui à identifier les multiples propriétés du Web dont certaines se révèlent surprenantes. L'efficacité vient du fait qu'elles épousent les contraintes des propriétés connues du Web comme les *patterns* que forment les liens hypertextes à grande échelle : c'est par l'étude des phénomènes d'agrégation de connexions que l'on identifie des groupes ou des communautés sur le Web et les réseaux sociaux. C'est en comparant les liens entre des pages web et leurs contenus (disons, les *keywords* auxquels on peut les résumer) que l'on arrive à isoler un silo d'informations pertinentes sur un sujet donné, cet « endroit » du réseau où les internautes contribuent à alimenter des connaissances collectives. Plus encore que l'information ou le document, c'est avec le lien hypertexte qu'il fallait commencer nos explorations, ce qui nous apparaissait comme la clef de voûte du système. Comme les nanomatériaux, il tisse sur différentes échelles, de la plus singulière à la plus vaste, des propriétés uniques et originales qui semblent s'emboîter pour donner naissance à des formes jusque-là insoupçonnées.

On comprend mieux alors tout l'intérêt porté au premier modèle du Web en *nœud papillon*, car il annonce ce que tout le monde attend : un gigantesque travail de recensement pour savoir ce qu'est donc devenu ce vaste système depuis la publication de son protocole en 1989. Ce modèle n'est pourtant pas le plan général qui permettrait de concentrer les propriétés génériques de cette architecture documentaire distribuée à grande échelle et de décrire les lois qui en gouvernent l'évolution. Pour y parvenir, il faut encore parcourir un long chemin. Mais il restera le premier modèle graphique présentant l'articulation logique de plusieurs propriétés statistiques du Web comme système de documents, enfin concentrées de façon cohérente.

Personne n'a encore les clefs globales du système. Chacun aimerait en détenir les plans, pas seulement par curiosité scientifique, mais aussi pour assurer l'efficacité d'un service offert aux réseaux sociaux d'acteurs, pour suivre des phénomènes informationnels propagatoires comme les *rumeurs* ou les *tendances*, ou encore lancer des opérations de détection et de *tracking* de certaines thématiques ou d'entités nommées – des personnes physiques ou morales. Posséder le plan global de la distribution complète des contenus dans la structure hyperliée du système nous ferait

passer de la toile indistincte à l'encyclopédie numérique universelle, contributive et raisonnée. Une façon de valider l'image développée au milieu des années 1990 d'une « bibliothèque universelle et ouverte à toutes les contributions ». Mais on est encore loin, à vrai dire, d'avoir isolé toutes les propriétés du Web et de pouvoir le penser *comme un tout*, comme un *système*. Des pans entiers de son architecture restent encore dans l'ombre et l'objectif de pouvoir les photographier est d'autant plus difficile qu'il faut accepter d'évoluer dans un univers qui ne semble admettre ni la clôture, ni l'exhaustivité, ni la classification hiérarchique ou arborescente. Le Web n'est pas contrôlable, du moins avec les instruments classiques de gestion des systèmes de connaissances.

On peut même se demander s'il s'agit tout simplement d'un *système d'information*. La question paraît toujours aussi curieuse pour un pur produit de la technologie numérique, peuplé d'autant de machines et d'ingénieurs, un artefact culturel nourri de bases de données, d'algorithmes et d'archives, de supports réductibles à une suite de 0 et de 1... et dont personne, en réalité, n'arrive à maîtriser réellement les masses, la diversité ou la dynamique évolutive. À bien des égards, le Web reste encore un mystère. Depuis le lancement de la première page web en 1989 par une équipe du CERN (Organisation européenne pour la recherche nucléaire), notre ignorance a grandi au fur et à mesure de son développement exponentiel, le Web conservant toujours un temps d'avance sur l'observation. Tim Berners-Lee, son inventeur officiel, a même lancé le projet d'en faire une science, les *web sciences*, en affirmant en 2006 : « Nous avons créé le Web, et nous avons pour devoir de le comprendre ». Depuis lors, nous vivons encore dans l'effervescence associée à l'arrivée d'un nouveau champ de recherche, un moment privilégié où une grande variété de disciplines viennent se confronter aux *web data*, le trésor de guerre des géants californiens de l'information. On a ainsi vu mobilisée la physique, pour étudier la logique des *flux* de liens et leurs concentrations comme des *masses*. La physique et la chimie pour essayer de comprendre les moments de mutation du système web – la *percolation* – puisqu'il ne semble pas évoluer de façon uniforme ou unilinéaire. La biologie pour le saisir comme un être vivant complexe, *auto-organisé*. On peut aussi le définir comme un *écosystème*, tout à la fois hiérarchisable sous forme d'arborescence (comme la pyramide des espèces) et décomposable en une multitude d'éléments interdépendants. On peut aussi demander à la sociologie d'expliquer les logiques sociales à l'œuvre derrière les phénomènes massifs de connexion entre des acteurs-producteurs d'informations via des documents interposés. On peut même faire appel à des réflexions plus abstraites sur le Web comme *système complexe* ou encore en faire l'objet de prédilection d'une science à venir sous la forme d'une sorte d'anthropologie du numérique et des réseaux.

Image 1.3

La première page web publiée à partir des protocoles sur lesquels a travaillé Tim Berners-Lee depuis 1989. Le Web y est défini comme un projet hypermédia ouvert et participatif : « The World Wide Web (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents ».

Le chemin que notre petite équipe a emprunté n'avait rien, quant à lui, d'un travail théorique et nous n'abordions pas le champ des *web data* pour valider des hypothèses préconçues. Nous avons adopté une démarche inductive en faisant le pari que les données ouvriraient sur des idées nouvelles, sans nous soucier du genre de recherche auquel nous appartenions. Officiellement, nous ne faisons donc *rien*, les programmes scientifiques officiels et la gestion des disciplines n'autorisant guère en France les aventures comme celle dans laquelle nous nous sommes lancés il y a plus de dix ans. Pour certains, il s'agissait seulement d'un bricolage ingénieux de code informatique pour « faire des cartes » ; pour nous, il s'agissait d'entrer de plain-pied dans une branche (seulement) de ce qui allait devenir les *data sciences*. À force de lectures, de discussions et de nombreuses expérimentations, le tableau général du Web a avancé par petites touches successives depuis une quinzaine d'années. J'ose espérer y avoir contribué.

Le schéma du Web en nœud papillon, comme d'autres travaux dont les nôtres, s'inscrit dans un espace borné intellectuellement, remplissant un vide entre deux repères communs qui font les deux limites du tableau. La première, c'est ce fameux nuage³

3. <http://noahveltman.com/internet-shape/>.

qui accompagne tant de documents scientifiques ou pédagogiques pour symboliser le Web dont le *cloud computing* est aujourd'hui l'écho.

Image 1.4

Représentation d'Internet ou du Web dans la littérature scientifique, d'après le travail de Noah Veltman, "What shape is the internet?" : <https://noahveltman.com/internet-shape/>.

Je n'ai jamais su s'il s'agissait ainsi de se dispenser des détails (on dessine le *nuage* comme on dit la « toile ») ou bien d'un aveu d'ignorance, symbolisant ainsi une sorte d'extériorité incompréhensible. L'autre limite, ou borne conceptuelle, est constituée par toute une lignée de travaux sur le *deep web*⁴, au départ nommé *invisible web* ou *hidden web*. On range dans cette catégorie l'ensemble des documents ou des informations inaccessibles aux moteurs de recherche et à leurs « robots » d'indexation en y mêlant les sites illégaux, les bases de documents formalisés (publications scientifiques, informations juridiques...), des sites gouvernementaux. Volontairement ou non, ces données à l'accès difficile semblent faiblement reliées au reste du Web, le Web de surface tel que nous le présentent les moteurs de recherche. On pourrait aussi y ranger le *dark web* autour duquel gravitent des sites de transactions commerciales illégales ou liés aux activités des *hackers*. Certains ont estimé qu'il pouvait abriter 90 % de l'information du réseau, mais accessible seulement à des profondeurs insondables, et seulement avec une logistique comme celle du projet Memex⁷ lancé en 2014 par la DARPA (Defense Advanced Research Projects Agency). Ces profondeurs extrêmes fascinent toujours, mais restent hors de portée des projets comme le nôtre.

4. "The Deep Web: Surfacing Hidden Value", *The Journal of Electronic Publishing*, 7 (1), juillet 2001, <http://www.press.umich.edu/jep/07-01/bergman.html>.

Image 1.5

Le Web se présente différemment selon sa profondeur, à mesure qu'on s'éloigne la surface, c'est-à-dire l'information présentée par les moteurs de recherche.

C'est là, entre ces deux extrêmes, le nuage et le *deep web*, que s'étend le Web tel que nous le connaissons et dont il s'agit de trouver le plan, le modèle, la carte. Pour l'instant, il ne s'agit que d'une esquisse, mais, en traduisant patiemment les masses de données en propriétés statistiques et visuelles, le Web commence à « prendre corps ».

Chapitre 2. Questions de distances

« Chatting to strangers becomes easier when based on a shared interest. One of the earliest uses of the Internet was to connect geographically scattered groups with shared minority interests. This “virtual communities” of people linked electronically may meet occasionally for what cyber-enthusiasts call “free-time” – or may never meet. Such *horizontal communities*, scattered around the world, may have more in common with one another than with their next-door neighbors », Frances Cairncross, *The Death of Distance: How the Communications Revolution Is Changing Our Lives*, 2001.

Au début de l'année 2014, des chercheurs allemands et italiens ont publié dans un article¹ une série de mesures récentes du Web pour vérifier certaines propriétés du « nœud papillon » exposées dans l'article de 2001, soit 13 années auparavant. Parmi ces mesures, en figure une qui ne laisse pas d'étonner au premier abord : la mesure du *diamètre du Web*. Il se situerait autour de 12 « clics » alors qu'il était de 16,3 dans l'étude de 2001 où apparaissait le modèle en nœud papillon. Qu'est-ce qui pouvait bien expliquer une telle différence ? Et, pour commencer, le réseau est-il donc réellement mesurable en termes de distance hypertexte de façon si certaine que l'on pouvait en tirer un diamètre ? Autrement dit, admet-il une *topologie*² ? La mise au point de métriques de distances pour un système documentaire comme le Web est toujours un sujet de discussion pour les chercheurs et les ingénieurs, habitués aux débats méthodologiques sur les procédés d'extraction, puis de traitement des données web. Pour un néophyte, ces questions très techniques de *distances entre des*

1. “Graph Structure in the Web – revisited: A Trick of the Heavy Tail”, par des chercheurs de l'université de Mannheim, <http://vigna.di.unimi.it/ftp/papers/GraphStructureRevisited.pdf>.

2. B.A. Huberman, *The Laws of the Web: Patterns in the Ecology of Information*, Cambridge (MA), MIT Press, 2001.

documents ne sont ni spontanées, ni même parfois concevables. Hormis pour des questions d'aménagement d'une bibliothèque, il ne viendrait à l'idée de personne de mesurer en centimètres l'espace qui sépare un ouvrage d'un autre.

Image 2.2

Le modèle du Web en nœud papillon. À gauche, d'après : A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins et J. Wiener, "Graph Structure in the Web", *Computer Networks*, 33 (1-6), 2000, pp. 309-320. À droite, d'après : R. Meusel, S. Vigna, O. Lehmborg et C. Bizer, "Graph Structure in the Web -- revisited: A Trick of the Heavy Tail", dans *Proceedings of the 23rd international conference on World Wide Web*, 2014, pp. 427-432.

Cependant, si cette mesure des distances est effectivement possible (calculable selon une *métrique*), il faudrait alors admettre que les univers numériques de l'information, et surtout le Web, ne font pas disparaître les distances, l'espace ou le principe de proximité : ils ne feraient, en résumé, qu'en réaménager la forme. Reste à mettre au point les instruments et leurs métriques, comme l'ont rapidement entrepris des chercheurs il y a une vingtaine d'années. Le Web peut donc avoir sa propre géographie, une sorte de *cybergéographie* du réseau. Et pour comprendre les méthodes utilisées depuis lors, il faut admettre deux principes.

Le premier consiste à supposer que les relations hypertextes constituent les unités de mesure d'un système qui est d'abord un *réseau*. Ce ne sont donc pas les masses d'informations ou les contenus documentaires qui permettent de le mesurer, mais l'estimation de leur degré d'interconnexion. Les premières mesures de distances sur le Web autour des années 1995 se sont donc appuyées sur la distribution des liens hypertextes entre les pages web. Il ne s'agit pas d'une abstraction sans fondement matériel puisque les liens hypertextes sont *écrits* dans le code HTML des pages web : ils sont donc à ce titre repérables et indexables dans des archives. Ces mesures tiennent compte d'une particularité du Web, et que l'on retrouve dans d'autres univers informationnels comme celui des publications scientifiques ou des brevets : quand une page web *A* pointe vers une page *B* (elle la *cite*), il n'est jamais assuré que *B* pointe en retour vers *A*.

Nous sommes donc entrés dans un espace qui n'est pas réversible, à moins de jouer sur notre mémoire de navigation avec la fonction « retour » du *browser web* préféré. C'est de cette particularité que naît l'une des propriétés du Web, parmi les plus intrigantes : lorsqu'on l'étudie à grande échelle, les liens hypertextes s'y donnent comme des *flux* dans un espace fondamentalement orienté, vectorisé comme par magnétisme. Le Web peut alors être saisi sous l'angle de sa physique, avec ses forces d'attraction et de répulsion dans un univers en continuelle expansion. La visualisation de ces phénomènes à toutes les échelles du Web laisse clairement apparaître des zones qui *diffusent* des flux de liens et d'autres, au contraire, qui semblent les *concentrer* – les attirer. Dans certains cas, on peut même constater qu'une seule page, sans produire de lien hypertexte, « tient » dans le réseau parce qu'une foule d'autres pages la pointent. C'est ce que l'on observe à petite échelle sur le Web (disons, quelques milliers de pages web) quand on cartographie les pages produites par des adeptes d'une secte qui pointent massivement vers celles du *leader* ou du guide. Il s'agit là d'une situation extrême qui peut indiquer, pour certains, qu'il existe une mesure possible de l'autorité (morale, intellectuelle, sociale...) dans des environnements distribués en dénombrant le nombre et la « qualité » des liens entrants dans une page ou URL.

Image 2.3

Cartographie des principaux sites mondiaux de diffusion de données scientifiques en 2008 (projet WebDatarium). La taille est fonction du nombre absolu de liens hypertextes qui relie une page au réseau ; la courbure des liens indique le sens de la connexion (de la source à la destination en tournant dans le sens des aiguilles d'une montre). L'architecture hypertexte du Web semble donc constituée par des phénomènes de densité de liens, mais aussi de direction des flux.

Prendre en compte l'orientation des liens ou des connexions pour construire un plan, une carte ou un modèle du Web n'est pas sans conséquence intellectuelle et technologique. Si la structure du Web est d'abord constituée matériellement des liens orientés, alors les instruments d'arpentage de ce nouveau territoire doivent en épouser le nombre et les orientations (pour tout *ratissier* – si l'on développe un moteur de recherche – ou simplement pour s'y *promener* – quand on est internaute). S'il n'existe pas de liens pour accéder à une page web, elle n'existe tout simplement pas. Transposé dans le contexte d'une bibliothèque, le principe reviendrait à considérer comme invisibles, inaccessibles (donc inexistantes) des ouvrages parce qu'ils n'en citeraient aucun autre ou ne seraient pas eux-mêmes cités ! C'est là tout le principe des moteurs de recherche : ils pointent pour nous ce qu'ils ont trouvé sur le Web en parcourant latéralement la structure étoilée. Ils peuvent aussi se servir de l'analyse de la direction des liens pour identifier tous les flux hypertextes. Il n'en faut pas plus pour inventer Google et le *pagerank*, son algorithme historique³. L'aspect le plus fascinant du principe du nombre et de l'orientation des liens entre documents consiste à lui associer le constat du développement exponentiel du Web : jour après jour, année après année, chaque nouvelle page diffusée vient alimenter des flux, et donc les renforcer, ou les modifier. Si l'on peut comparer ces flux orientés à des forces d'attraction qui ouvrent sur un univers dont on constate l'*expansion continue* en nombre de documents, alors on en vient naturellement à emprunter aux astrophysiciens leurs modèles d'un univers en *expansion* dont les forces locales de *gravitation* ne font que se renforcer dans le même mouvement. Surprenante, cette façon de considérer le Web rappelle combien les recherches exploratoires – risquées pour le chercheur ou l'ingénieur – se rapprochent des meilleures œuvres de science-fiction, un réservoir inépuisable de scénarios ou de décors ! On pourrait prendre l'idée pour une spéculation théorique... jusqu'au jour où, à l'occasion d'un congrès de mathématiques en 2003, trois chercheurs français en valident à leur tour le principe⁴ en proposant un *modèle gravitationnel du Web*.

Le second principe qui fonde la capacité à mesurer le Web à titre d'espace relationnel tient en un mot : le *graphe*. Le principe en paraît simple : en projetant des éléments (*nodes*) dans un espace et en leur associant des liens (*links* – éventuellement orientés), on peut alors calculer nombre de propriétés de la structure produite. La théorie des graphes, fondamentalement, est un instrument de réduction logique de la complexité et ses métriques associées constituent une sorte de « boîte à outils » nécessaire

3. S. Brin, L. Page, R. Motwani et T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford Digital Library Technologies, 1999 ; S. Brin, L. Page, R. Motwani et T. Winograd, "What can you do with a Web in your Pocket?", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 21 (2), 1998.

4. T. Bennouas, M. Bouklit et F. de Montgolfier, *Un modèle gravitationnel du Web*, Montpellier, LIRMM, 2003.

à toute exploration du Web. Techniquement, elle repose sur la construction d'une table des croisements ou une matrice des connexions (ou *matrice d'adjacence*) entre des éléments que l'on sait dénombrer (éléments discrets). Quand on les découvre pour la première fois, ces matrices de graphes font parfois penser aux tables de nos atlas routiers avec les villes en abscisse et en ordonnée, à la différence que ce qui est indiqué dans notre tableau, ce ne sont pas les distances kilométriques, mais la présence ou l'absence d'une connexion. Le succès de cette méthode de modélisation ou de description de phénomènes complexes comme les réseaux de distribution de l'énergie, les flux financiers mondiaux, la propagation des virus ou les circuits de l'eau dans l'environnement vient de sa souplesse inhérente. Par exemple, dans nos atlas routiers, figure toujours une diagonale où l'on comprend que chaque ville, à l'endroit où se croisent son abscisse et son ordonnée, ne peut avoir de distance avec elle-même. Mais on peut au contraire, dans le cas des matrices de données web, indiquer la connexion d'un élément avec lui-même (les liens internes des sites web entre les pages), accorder des poids aux liens avec une valeur numérique (la valeur des échanges bancaires entre deux comptes) ou encore associer deux nœuds avec plusieurs liens (les multiples formes d'interaction sociale entre deux acteurs). Et l'on peut aller encore plus loin : sur nos atlas routiers, les distances entre deux villes sont équivalentes quel que soit le sens du trajet. Pour le trajet Paris-Lille comme pour le trajet Lille-Paris, la distance est la même, donc dans l'atlas routier on retrouvera le même chiffre deux fois, et donc de part et d'autre de la diagonale où les villes se croisent elles-mêmes. Sur le Web, il en va tout autrement : si une page est pointée par une autre, la réciproque n'est pas forcément vraie. Ainsi, les matrices d'adjacence qui permettent d'étudier la distribution des liens hypertextes entre des pages ou des documents sur le Web ne contiennent pas les mêmes informations de chaque côté de la diagonale – elle n'est pas symétrique.

Image 2.4

Différents types de réseaux : simple, avec des autoliens, des liens pondérés, des liens parallèles, et des liens orientés.

Appliquée à grande échelle, cette puissance de réduction des phénomènes complexes ouvre rapidement sur des calculs d'une grande complexité, et qui peuvent durer un temps certain pour nos machines. À supposer que l'on ait la capacité d'arrêter le Web à un instant t pour le photographier entièrement, la matrice des pages comporterait plusieurs dizaines de milliards d'éléments en abscisse et en

ordonnée avec plusieurs centaines de milliards de 1 là où deux pages sont reliées et de 0 là où elles ne le sont pas. En théorie, une carte du Web est donc possible et nombreux sont ceux qui aimeraient explorer les propriétés d'un tel graphe, à supposer que l'on ait les moyens techniques, humains et financiers de le produire... et de le partager. Car les « géants » de l'industrie de l'information web possèdent probablement des graphes complets et régulièrement remis à jour de leurs périmètres d'intervention sur le Web : Google pour déployer son algorithme PageRank, LinkedIn et Facebook pour identifier les relations possibles de profil à profil ou de compte à compte, Twitter pour analyser la dynamique des marqueurs que sont les *hashtags*. À défaut, on pourra tester des extractions massives (mais partielles) de la structure hyperliée du Web sur des jeux de données accessibles publiquement, comme l'ont fait les chercheurs allemands et italiens dans leur vérification récente des propriétés du modèle du Web en nœud papillon⁵.

Ce sont ces grands graphes qui ont rendu le Web mesurable, à partir d'une batterie de métriques qui laissent en apercevoir quelques propriétés essentielles comme son *diamètre* ou sa *densité*. En quelque sorte, la théorie des graphes assure la transformation des *web data* en qualités physiques presque palpables, visuelles. Si l'on reprend la première étude du Web comme un nœud papillon en 2001, le nombre moyen de liens hypertextes par page est autour de 7 alors qu'il est dans l'étude des chercheurs allemands et italiens de plus de 35. Comment expliquer un tel écart entre le Web de 1998 et celui de 2013 ? C'est peut-être le rôle que jouent aujourd'hui les feuilles de style CSS et le HTML5 qui densifient la structure hyperliée des sites web, et donc la richesse des parcours possibles de navigation. D'autres facteurs peuvent expliquer ce phénomène d'augmentation de la connectivité du réseau, mais il semblerait qu'à grande échelle les pages sont prises dans une structure qui se densifie au cours du temps alors que le Web s'est exponentiellement développé... Voilà une question qui ne laisse pas d'intriguer, tout comme celle du diamètre d'un système aussi vaste et dynamique. En théorie des graphes, on base l'estimation d'un système en termes de grandeur à partir du calcul des chemins possibles dans la structure

5. Parmi les différents sets de données massives sur le Web accessibles publiquement : l'*open repository of web crawl data* (<http://commoncrawl.org/blog/>) extrait du Web en août 2004 avec 3,5 milliards de pages et 128 milliards de liens hypertextes, le WebScope de Yahoo!Labs (<http://webscope.sandbox.yahoo.com/>), le ClueWeb12 (<http://www.lemurproject.org/clueweb12.php/>) avec plus d'un milliard de pages. Parmi d'autres sets accessibles et plus variés, le Stanford Large Network Dataset Collection (Jure Leskovec, <http://snap.stanford.edu/data/>) et la page de datasets de M.E.J. Newman (<http://www-personal.umich.edu/~mejn/netdata/>).

hyperliée. Pour cela, il faut reprendre la matrice et dénombrer pour chaque paire de nœuds (ou chaque couple possible de pages web) chacun des chemins possibles (ou non), et les classer du plus court au plus long. Si on réalise le calcul pour chaque paire de nœud (ou alors un échantillon), on peut extraire des informations déterminantes pour saisir les propriétés physiques essentielles du Web : on peut ainsi estimer le diamètre d'une structure construite sur des connexions en réalisant la moyenne des plus longs chemins possibles entre chaque paire de nœuds reliés. On peut aussi réaliser la même moyenne, mais, à partir des chemins possibles les plus courts, c'est le rayon. Et, de façon plus subtile, identifier les nœuds par lesquels passent le plus grand nombre de chemins possibles, l'une des façons d'estimer le degré de *centralité* d'un nœud (donc d'une page sur le Web). Ce type de mesures permet souvent de classer les éléments d'une structure en fonction de leur rôle organique, comme des pierres de voûte dans l'architecture documentaire. Au-delà de l'ingénierie des données et des expérimentations scientifiques, la notion de centralité, quand elle est appliquée à des réseaux sociaux (sur le Web ou non) a rapidement été appliquée à de nombreux métiers depuis les années 2000 : la recherche des *influenceurs* en marketing, les *relais* d'opinion en communication ou les *innovateurs* dans un écosystème économique.

À partir de ce type de méthode, il est donc possible d'établir le diamètre du Web : *16,12* dans la première étude de 2001 ; *12,84* dans la seconde, treize ans plus tard. Si l'on tient compte du fait que le graphe de la seconde étude comporte près de quinze fois plus de nœuds, il apparaît rapidement que le diamètre du Web *diminue malgré son développement exponentiel* durant ces années⁶. Un développement massif de la connectivité semble accompagner son extension, et donc le nombre de pages publiées. Beaucoup plus densément peuplé de documents et de liens en 2013 qu'en 2001, le Web semble se concentrer et offrir de nombreuses possibilités de le parcourir rapidement. Son espace se rétracte, et le temps s'accélère sur le réseau.

Ce principe de conservation, voire de diminution des distances malgré la multiplication des éléments se vérifie aussi ailleurs, notamment sur les réseaux sociaux en ligne. Ce sont des mondes circonscrits sur le Web et la plupart du temps liés à des acteurs privés. On peut donc les décrire de façon plus certaine et plus exhaustive que le Web en tant que tel, et nombre de géants californiens en tirent quelques leçons intéressantes pour leurs activités marketing. Ainsi, il y a quelque temps, la publication d'une enquête réalisée par Lars Backstrom et des chercheurs de l'université de Milan a fait grand bruit sur le Web : la distance moyenne (calculée en nombre d'amis communs) entre deux comptes Facebook serait aujourd'hui de 4,7 au lieu

6. J. Leskovec, J. Kleinberg et C. Faloutsos, "Graph Evolution: Densification and Shrinking Diameters", *ACM Transactions on Knowledge Discovery from Data*, 2007.

de 5,2 en 2008. À ce jour, il s'agit du plus grand graphe social jamais construit avec 721 millions de nœuds (comptes Facebook) et plus de 69 milliards de liens, chaque utilisateur de la plateforme ayant en moyenne 190 « amis ». Cela en fait un graphe qui paraît relativement dense et c'est évidemment beaucoup plus que la moyenne de 7 liens hypertextes par page web en 1999 et de 35 en 2014 des deux études sur le *modèle en nœud papillon*.

L'important, ici, est de noter que cette distance semble diminuer au cours du temps, à la mesure de la richesse et de la diversité des connexions sociales produites sur la plateforme (phénomène vérifiable aussi bien à l'échelle mondiale qu'à l'échelle locale des États-Unis). Il s'agit donc d'un phénomène similaire à celui observé sur le Web de façon générale : les univers informationnels se contractent à mesure qu'y entrent de nouveaux éléments. Les deux études sur le modèle du Web en nœud papillon et celle sur Facebook peuvent être rapprochées si l'on envisage que les plateformes sociales ont fondamentalement pour objectif d'accélérer le processus d'interconnexion entre des acteurs qui y ouvrent des comptes (techniquement, des pages web qui contiennent des informations personnelles). Leur rôle est aujourd'hui fondamental dans l'écosystème d'information que constitue le Web. Elles couvrent la plupart des aspects de la vie sociale en ligne – surtout si l'on inclut les plateformes de jeux – et leurs créateurs aiment souvent à rappeler l'ensemble des niches d'activités qu'elles abritent et dont elles se nourrissent (« MySpace, c'est le café, Facebook, le barbecue au fond du jardin, et LinkedIn, le bureau », résume Reid Hoffman, le créateur de LinkedIn). Ces technologies de la connexion constituent la colonne vertébrale de ce second âge du réseau, le Web 2.0. On peut aussi le dire autrement : si le Web est techniquement construit sur des liens hypertextes qui relient des documents et des contenus entre eux, il est surtout alimenté par la dynamique sociale des internautes. Et c'est toute la perspective de l'étude de Lars Backstrom et de ses collègues qui démontre combien, derrière la mesure d'un diamètre aussi réduit pour une aussi grande masse de comptes, Facebook permettrait donc de réduire les distances sociales, géographiques et culturelles et, en quelque sorte, de rapprocher les hommes d'eux-mêmes. Cet argument, très « marketé » par Facebook ou d'autres acteurs, illustre à sa façon combien le développement des technologies réseau s'est toujours accompagné d'un rêve d'universalité, de réunion humaniste entraînant dans son sillage toutes les figures d'une proximité sociale inédite : le marketing d'un réseau social planétaire où tout le monde dans sa diversité se trouve enfin connecté à lui-même, des relations symboliques fraternelles, avec le Dalaï-Lama ou Barack Obama, des mouvements révolutionnaires menés par des jeunes connectés au Web, la force des mobilisations collectives via les réseaux (les OGM, le réchauffement climatique, les *parabens*...), le partage de connaissances ouvertes à travers Wikipedia. Comme le souligne l'auteur : « Thus,

when considering even the most distant Facebook user in the Siberian tundra or the Peruvian rainforest, a friend of your friend probably knows a friend of their friend ».

4,7 aujourd'hui sur Facebook, 5,2 en 2008 : les chiffres font rêver et, effectivement, ils doivent être pris pour la signature statistique d'un *petit monde*. Ils s'inscrivent dans la (longue) lignée des travaux expérimentaux sur la forme ou la « géographie » des réseaux sociaux, dont le champ s'est considérablement élargi et renouvelé depuis l'arrivée des technologies numériques et les calculs menés à grande échelle sur les données d'usage. Au-delà du texte visionnaire de Karinthy (*ChainLinks* en 1929⁷), c'est l'expérience de Stanley Milgram⁸ avec le « jeu des lettres » qui va asseoir pour la postérité le fameux chiffre « 6 » (ou *six degrés de distance*) qui reste encore aujourd'hui une sorte de repère symbolique pour tous ceux qui cherchent à comprendre la topologie des réseaux sociaux, et en particulier l'une de leurs principales propriétés statistiques : les distances moyennes très réduites entre chacun des nombreux acteurs d'un système en termes de « contacts » ou de « connexions ». L'étude de Lars Backstrom n'est que le jalon le plus récent d'une série continue de travaux sur l'*interconnected world* qui paraît de plus en plus *small* au fur et à mesure que le Web étend sa toile : Microsoft le démontrait déjà en 2008 dans l'analyse des données associées à Messenger parmi tant d'autres et, pour aller au-delà, chacun peut aujourd'hui mesurer en ligne les distances entre acteurs de cinéma jouant dans les mêmes films avec l'Oracle de Bacon⁹ ou encore les expérimentations sur la distribution des liens entre articles de Wikipedia.

Et puis il y a l'ouvrage légendaire de Duncan Watts, *Six Degrees: The Science of a Connected Age*. Publié en 2004, *Six Degrees* a marqué les esprits pour son caractère démonstratif de certaines des lois qui gouvernent des systèmes comme le Web, en particulier celle qui concerne les distances étonnamment courtes pour des univers qui peuvent compter jusqu'à plusieurs dizaines de milliards d'éléments (et parfois doivent se synchroniser comme le rappelle son collègue Steven Strogatz – *Sync. The Emerging Science of Spontaneous Order*, 2003). Pour sa démonstration, Duncan Watts se sert d'une sorte de « tuner analytique » sur lequel il joue et qui comporte deux extrêmes théoriques et expérimentaux. Le premier extrême est celui d'un modèle de graphe régulier, presque strict dans son ordonnancement où tous les nœuds ont un même nombre de liens (par exemple 4) qui pointe vers 4 voisins immédiats. Ce modèle, d'ordre absolu, pourrait être visualisé comme les mailles d'un filet de pêche et où chaque nouveau nœud et ses 4 liens viendraient se

7. https://djjr-courses.wdfiles.com/local-files/soc180%3Akarinthy-chain-links/Karinthy-Chain-Links_1929.pdf.

8. J. Travers et S. Milgram, "An Experimental Study of the Small World Problem", *Sociometry*, 32 (4), 1969, pp. 425-443.

9. <https://oracleofbacon.org/>.

positionner en périphérie, agrandissant ainsi la surface du filet. Ce modèle logique construit sur la régularité (4 liens pointant sur des voisins immédiats ou contigus) génère des distances énormes en termes de parcours et un nombre presque infini de chemins. Nous sommes donc très loin des distances courtes observées sur le Web ou les réseaux sociaux. Le second extrême est constitué par un modèle de graphe qu'aiment à produire les mathématiciens ou les physiciens pour identifier certains aspects des systèmes complexes : les structures aléatoires ou *randomness*. Pour en saisir le principe, il faut s'imaginer tous les nœuds d'un système entre lesquels on distribuerait le nombre et la direction des liens au hasard. N'importe quelle formule générant un principe aléatoire suffit, par exemple un jeu de dés : prenez toutes les paires possibles de nœuds et, pour chacune d'elles, lancez les dés. Si un 6 apparaît, il y a alors un lien entre un élément *A* et un élément *B* et dans n'importe quel autre cas (si vous tirez 1, 2, 3, 4 et 5, il n'y a pas de lien entre *A* et *B*). On peut faire tourner un tel type d'algorithme par passes successives pour peupler le graphe de liens, sans coût technologique important puisque l'expérience n'est pas générée à partir de données réelles, mais d'un modèle abstrait, artificiel. Dans ce cas, les distances sont courtes, voire très courtes en fonction du paramétrage du modèle, mais le graphe ne comporte aucune régularité, hormis quelques lois statistiques qui permettent d'entrevoir l'évolution du système.

Le Web, avec son diamètre court, n'est ni l'un ni l'autre de ces deux extrêmes : pour Duncan Watts, il est *entre les deux*. Dans sa démonstration, Watts se saisit d'un graphe régulier (le filet de pêche avec ses distances très longues) et y introduit un lien aléatoire, puis deux, puis trois... divisant ainsi par 50 le diamètre de la structure à chaque fois ! Quelques liens aléatoires jetés sur une structure régulière et fonctionnant par contiguïté suffisent donc pour approcher un modèle cohérent du Web et compatible avec les données issues des campagnes de *crawl* de 1999 et 2013. Le tuner analytique fonctionne et l'auteur joue ainsi avec différents modèles générés à partir d'un curseur que l'on déplacerait entre *ordre* parfait et *chaos*, entre nécessité et hasard. Ou le 1 et le 0 d'une variable, calée pour le Web autour de 0,5.

Image 2.5

Les réseaux *small world* tels que conceptualisés par Duncan Watts dans son livre *Six Degrees*.

Ces liens aléatoires, jetés d'un bout à l'autre du graphe, faisant ainsi chuter son diamètre, sont déjà connus, et depuis longtemps, dans la sociologie ou plutôt la sociographie des acteurs du monde social avec Mark Granovetter¹⁰. Chez ce dernier, les *weak ties* (l'équivalent des liens aléatoires de Duncan Watts) tirent leur force dans l'étude de la macrostructure sociale de leur pouvoir de connexion de différents univers sociaux : tandis que les *strong* (local) *ties* (l'équivalent du principe de contiguïté dans le filet de pêche) permettent de décrire les clans familiaux et les univers d'amis proches, les *weak* (global) *ties* semblent jouer un rôle déterminant dans notre faculté d'adaptation sociétale au sens large (ils sont activés, par exemple, dans des périodes de recherche d'emploi chez les jeunes diplômés). De là, l'idée de la *force des liens faibles* qui assurent à une structure un diamètre relativement court malgré un nombre massif d'éléments. L'effet *small world*.

Reste à déterminer d'où viennent ces liens transversaux, qui ou quoi génère ces *short-cuts* ? Constatée localement en sociologie, validée théoriquement avec Duncan Watts, cette faculté d'adaptation des acteurs (en l'occurrence sociaux) peut en laisser plus d'un admiratif de la plasticité de l'espace social, de son caractère organique et dynamique. Et le Web et les réseaux sociaux permettent d'en vérifier le principe, et même de l'accélérer comme le prétendent les chercheurs de Facebook. Mais ces questions fascinantes de distance dans les réseaux en cachent d'autres, car, sur les réseaux sociaux, et peut-être à l'échelle globale du Web lui-même, l'effet *small world* se double d'un autre phénomène qu'un modèle potentiel du Web doit pouvoir

10. M.S. Granovetter, "The Strength of Weak Ties", *American Journal of Sociology*, 78 (6), 1973, pp. 1360-1380.

intégrer. *Clusters*, communautés, groupes, tribus, clics, classes, différences... : à chaque fois que l'on étudie de près des données web, le calcul des distances étonnamment courtes laisse aussi apercevoir des univers séparés, cloisonnés, dispersés en une myriade d'éléments agrégés ensemble, mais séparés des autres, quasiment autonomes pour certains. Et ces particularités locales (on pourrait dire des *gravités communautaires*) augmentent peut-être elles aussi avec le temps, à l'image de ce qui se passe à grande échelle sur le Web avec les forces d'attraction grandissantes d'un côté et l'expansion continue de l'autre. Il faut parcourir l'ensemble de l'étude de Lars Backstrom et de ses collègues sur les comptes Facebook pour comprendre qu'une forme de topologie de l'agrégation sociale accompagne inmanquablement un diamètre court et qu'elle règle l'essentiel des profils de la plateforme sociale : 84 % des connexions relient des comptes d'un même pays, en général entre des abonnés du même âge et plutôt des voisins géographiques. Comment donc peut-on tous être à la fois si *proches* et si *séparés* sur le Web ?

La question résonne bien au-delà des limites de la sociologie, car les données issues des réseaux sociaux peuvent exhiber des *patterns* riches d'enseignements pour comprendre le Web lui-même. C'est très exactement le cas sur les questions de mesure de distances dans les réseaux d'information, un terrain merveilleux d'expérimentation technique. Et inversement, certaines propriétés statistiques posent inmanquablement des questions à nos façons d'envisager le Web sous l'angle social ou culturel¹¹. Sur ce point, je partage souvent l'enthousiasme de ceux qui vantent les bienfaits des technologies ouvertes et distribuées ou de ceux qui caressent l'espoir d'une humanité connectée en temps réel, comme libérée des contraintes géographiques ou sociales. Mais, plus on pousse l'analyse des données web réelles, à petite comme à grande échelle, plus l'univers social sur lequel ouvrent les calculs apparaît fortement clusterisé, morcelé, cloisonné. L'expansion des univers numériques semble ainsi se conjuguer inmanquablement avec l'augmentation des gravités locales. Si, à propos de Facebook ou d'autres espaces présentés comme issus du Web 2.0, on avance souvent les chiffres d'une proximité réduite entre les internautes, on porte beaucoup moins d'attention (y compris scientifiquement) à tous ces phénomènes, ces tendances, ces mouvements de discordes (la blogosphère au moment des élections politiques), de polémiques (controverse sur les OGM, les additifs alimentaires, l'Europe...), d'affrontements (le conflit israélo-arabe se déroule aussi sur la toile), de séparation et d'éloignement entre les acteurs. Dans l'*interconnected world*, la multiplication des liens transversaux peut aussi être l'occasion de la multiplication des conflits. Souvent, le Web est même un champ de bataille. Se déplacer rapidement d'un bout à l'autre de la structure web, en quelques « clics » seulement, c'est aussi survoler ces puissants attracteurs

11. M. Castells, *La galaxie Internet*, Paris, Fayard, 2002.

que sont les cultures et les langues et les cercles proches où s'agrègent les amis, la famille, la classe d'âge, la profession, les collègues, la religion, les opinions politiques... comme autant de critères de séparation, d'éloignement, de distance difficilement réductibles. 6, 5,2 ou 4,7 représentent des distances incommensurables en termes d'univers culturels, linguistiques, historiques à « traverser » (et selon quels critères ?). Facebook, comme les autres plateformes, se nourrit d'ailleurs principalement de ces proximités natives (à combien « d'amis » s'est-on lié avec les suggestions de type « Les connaissez-vous ? »). À l'inverse, combien de fois les internautes ont-ils réellement navigué dans ces blogs étranges, qui parlent de sujets curieux et exotiques, éventuellement dans une autre langue, à l'autre bout, non pas du monde, mais du graphe des 271 millions de comptes et des 69 milliards de liens les unissant sur Facebook ? Six degrés de distance, mais de *proximité* ou d'*éloignement* ?

Chapitre 3. Hiérarchies

« The Web has become a veritable laboratory, where one can study human behavior with a precision and on a scale never possible before. These studies are closer to social science than computer technology, and they profit from the fact that the reach and structural complexity of the Web creates an *Ecology of Knowledge*. This ecology is characterized by relationships, information “food chain”, and dynamic interactions that could soon become as rich as, if not richer than, many natural ecosystems », Bernardo A. Huberman, *The Laws of the Web: Patterns in the Ecology of Information*, 2001.

J'ai ouvert mon blog en 2011, l'Atelier de Cartographie. Je me souviens bien de ce moment et du travail consciencieux auquel je me suis astreint : la rédaction du premier post, puis de tous ceux qui sont venus s'accumuler chronologiquement par la suite, la découverte des fonctionnalités de Wordpress qui rendent possibles la publication de mes documents et l'insertion d'images ou de vidéos YouTube, la gestion des rubriques pour laquelle je me suis reposé sur un modèle gratuit de mise en pages. Me voilà donc, depuis plus de quatre années, modeste contributeur du plus vaste système documentaire jamais connu. Mais je me souviens aussi des efforts qu'il m'a fallu déployer pour sélectionner une série de « références » à faire figurer dans le *blogroll*, à droite des textes publiés. Tous les blogueurs, comme moi, y placent des « amis » proches de leurs préoccupations, des « pairs » ou des « semblables » avec lesquels nous nous agrégeons spontanément. Dans l'Atelier de Cartographie, j'y fais figurer des liens hypertextes qui renvoient vers des projets dont je me sens proche du point de vue personnel et qui comptent dans l'environnement d'innovation technologique où je navigue : Blue Node, Data Publica, Gephi, l'Atelier Iceberg, Linkfluence, Linkurious, Opendatasoft, Perfect Memory, READi, virtual sensitive. Personnellement, je reste toujours aussi surpris de voir un blogueur inscrire ainsi avec autant de certitude et précision son blog (ou son objet documentaire) dans l'espace du Web : les liens se distribuent avec minutie, en direction de tous les éléments qui

le situent dans un *voisinage*, une sorte de *proximité* topologique qui coïncide avec la *similarité* (relative) des contenus ou des thèmes.

Image 3.2

Mais il y a plus : figés depuis 2011 dans le *blogroll* de l'Atelier de Cartographie, figurent des liens vers ceux qui m'apparaissent comme des autorités scientifiques (Barabási, Watts, Kleinberg, Newman, Strogatz pour la théorie des graphes et les *network sciences*, mais aussi Places & Spaces, Mapping Science, des *authorities* dans le domaine de la cartographie de l'information scientifique et technique). Ici, la grande différence, c'est que ces objets documentaires *ne* reconnaissent *pas* mon blog, *ne* le pointent *pas*. Dommage, car j'aurais peut-être ainsi récupéré un peu de leur notoriété au passage d'une citation... À l'échelle du Web, l'univers de la cartographie d'information et des sciences des réseaux dans lequel j'ai placé mon blog ne constitue qu'une partie infime du système, comme une sorte de périphérie lointaine, dans une communauté relativement circonscrite qui intéresse quelques spécialistes comme moi ou des étudiants de passage. Cette localité thématique est à peine saisissable à l'échelle du Web, mais, en y braquant son objectif, le zoom fait vite apparaître une communauté riche et active où nous sommes nombreux à produire des contenus. Elle exhibe quelques aspérités statistiques remarquables, notamment ces sites prestigieux qui semblent se nourrir de liens entrants (Watts, Barabási) que tout le monde contribue à produire. Y compris moi-même depuis 2011, et avec application.

Une étude attentive de toutes les actions menées par un nouvel entrant dans un système comme le Web peut être tout aussi instructive que des études massives, notamment sur les stratégies de connexion qu'il déploie. N'ayant ni centre, ni gouvernement, ni système global de classement, le Web condamne inmanquablement l'utilisateur à déployer des stratégies locales de diffusion de ses documents comme le ferait

une espèce dans un biotope. La comparaison entre un écosystème documentaire contributif et distribué avec un écosystème naturel n'est pas une métaphore de circonstance ou un procédé pédagogique : pour les sciences du Web comme les sciences de l'environnement ou la biologie, un *écosystème* ne se réduit pas à la somme de ses éléments dénombrables, voire aux connaissances accumulées à leur sujet. Il s'agit, dans les deux cas, de comprendre la nature des interactions multiples générées par leurs éléments selon différentes règles ou lois, et à différentes échelles. De ces innombrables interactions émergent des formes ou des *patterns*, l'un des principes fondateurs du concept de biodiversité. Depuis les années 1997-1998, les ingénieurs et les chercheurs ont rapidement compris qu'il en allait de même avec le Web comme *biotope documentaire* que l'on peut étudier avec des opérations massives d'extraction de données (d'où naît le *modèle en nœud papillon*), mais aussi en prêtant une grande attention au comportement d'un seul élément entrant dans le système. Les stratégies d'un nouveau contributeur, son comportement, les actions avec lesquelles il noue des relations proches ou lointaines avec ses semblables (donc compatibles techniquement) peuvent expliquer la façon dont il renforce par sa seule présence au Web des tendances statistiques ou des *patterns* déjà massivement présents. Ainsi, quand j'ouvre mon blog en 2011, je m'inspire de « ce que c'est qu'un blog et comment ça marche » en suivant des exemples déjà existants ou des modèles proposés par Wordpress. Un exemple de ce que Jon Kleinberg va appeler en 1999 une *copying strategy*¹ et qui se vérifie également à très grande échelle sur le Web. Mais depuis 2011, aussi, je prends soin de pointer dans mon *blogroll* comme dans mes *posts* ces autorités scientifiques qui rayonnent dans mon domaine et auxquelles, en quelque sorte, je paie spontanément un tribut. Un exemple de ce que Albert-László Barabási et Réka Albert vont appeler en 1998 un principe de *preferential attachment*² : le nouvel entrant dans le système cherche à se connecter aussi bien à ses proches qu'aux autorités centrales de référence dans un domaine. Connexions *latérales* vers ses semblables (souvent réciproques) ou ses « proches » et toute la dynamique d'interaction comme dans les connexions entre espèces dans un biotope ; connexion *centrale* vers un cœur-attracteur, faisant irrémédiablement penser aux espèces clés au sommet de la pyramide de classification du vivant : la balance est parfaite.

Reste évidemment à articuler ces deux principes, ces deux lois. Ensemble, elles se fondent sur des usages « programmés » qui assurent la *prédictibilité* de la distribution des liens dans un agrégat (ou un « cluster d'amis ou de proches ») ou à l'échelle du Web. Si l'on peut deviner la façon dont les liens vont apparaître, on devrait donc

1. J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan et A.S. Tomkins, "The Web as a Graph: Measurements, Models, and Methods", dans *COCOON'99: Proceedings of the 5th annual international conference on Computing and combinatorics*, 1999.

2. A.-L. Barabási et R. Albert, "Emergence of Scaling in Random Networks", *Science*, 286 (5439), 1999, pp. 509-512.

maintenant pouvoir mieux comprendre l'organisation générale du système entier, et donc aussi son développement à venir... La forme du Web *et* son histoire. Et la plus évidente propriété du Web qui semble se développer, c'est sa nature profondément *hiérarchisée*. Si l'on regarde attentivement, non pas le *nombre moyen* de liens hypertextes par page ou par documents, mais la nature de leur *distribution* à grande échelle, une courbe magnifique apparaît et croît au cours du temps, comme le prévoient les modèles théoriques, mais aussi les observations. Une courbe en *loi de puissance*, une *power law*. Si l'on met en abscisse chacune des pages du Web (ou d'un corpus de pages web) et en ordonnée leurs nombres réels de liens hypertextes, un phénomène massif de hiérarchisation semble régler la connectivité du réseau, la fameuse loi de puissance, bien connue des économistes ou des linguistes, et depuis fort longtemps. De façon générale, cette courbe indique que dans l'univers exploré une minorité d'éléments monopolisent la grande majorité des liens. Et inversement, évidemment : une grande majorité d'éléments semblent se partager une minorité de liens. C'est la loi bien connue des 80/20 : 20 % des pages web accumulent 80 % des liens et 20 % des liens hypertextes sont distribués parmi 80 % des pages restantes d'un jeu de données.

Image 3.3

À gauche : distribution normale. Un petit nombre de sites ont peu ou pas de liens pointant vers eux, un grand nombre de sites en a un nombre modéré, et peu de sites en ont un grand nombre. À droite : distribution en loi de puissance. Un petit nombre de sites a énormément de liens pointant vers eux, tandis que la plupart n'en ont pas ou peu.

Cette courbe de distribution peut varier selon les données web analysées. Dans certains cas, le phénomène peut être encore plus hiérarchisé avec une loi de type 10/90 et la courbe tend vers une sorte d'angle droit. Forcez le phénomène, et vous obtiendrez un univers où tous les éléments ne sont reliés qu'à un seul d'entre eux, l'Unique qu'aimeraient incarner Google, Twitter ou Facebook. Dans d'autres cas, le phénomène est moins accentué, laissant apparaître une organisation hiérarchisée plus diluée avec une courbe plus arrondie. Dans tous les cas, le phénomène se vérifie sur

des grandes masses de données web. Si l'on n'a pas oublié que les liens hypertextes sont orientés, le phénomène de hiérarchisation peut être analysé du point de vue des liens *entrants* comme des liens *sortants* des nœuds d'un graphe.

On peut en raffiner la description et entrer dans les mathématiques propres aux réseaux à invariance d'échelle (*scale-free networks*) dont Albert-László Barabási a posé les fondements à partir du calcul d'un coefficient ou invariant, dans *Linked*, un autre ouvrage légendaire des *network sciences*³.

Image 3.4

Les réseaux *scale-free* conceptualisés par Barabási dans *Linked*.

Si l'on doit la topologie du Web à la façon dont les liens se distribuent entre les éléments ou les *nœuds* d'un système, alors apparaît clairement le rôle majeur que jouent les *hubs* vers lesquels se concentre la connectivité. À vrai dire, ils sont indispensables au système et constituent des aspérités statistiques remarquables, comme les massifs en géographie. Stanley Milgram, dans sa célèbre étude autour du « jeu des lettres », avait déjà remarqué à la fin des années 1960⁴ le rôle majeur que jouent les *socio-stars* dans les chemins empruntés par les courriers postaux à travers les États-Unis. On fait remonter la découverte de cette loi statistique de distribution à l'économiste italien Vilfredo Pareto qui, en étudiant les données fiscales de différents pays, puis certains aspects de l'entreprise sous l'angle des finances ou de la gestion des stocks, en avait remarqué le principe. Si l'on réduit le champ d'observation du Web aux plateformes contributives sur le réseau, le même phénomène de hiérarchisation gouverne l'activité des contributeurs (10 % sont très actifs tandis que 90 % contribuent faiblement ou occasionnellement). Si l'on intègre les utilisateurs, les chiffres s'enchaînent comme dans un jeu de poupées russes : 1 % de contributeurs actifs, 9 % de contributeurs occasionnels et 90 % d'utilisateurs non-contributeurs. Un examen d'une plateforme comme GitHub illustrerait le même principe, aussi bien du

3. A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, New York, Basic Books, 2002.

4. J. Travers et S. Milgram, "An Experimental Study of the Small World Problem", *Sociometry*, 32 (4), 1969, pp. 425-443.

côté contributeurs que du côté des projets qui y sont déposés. Idem sur Steam où les *gamers* comme les jeux sont distribués en *power law*.

Image 3.5

Les principales propriétés des distributions en loi de puissance.

Ce principe peut paraître fascinant, presque fractal, surtout si l'on est penché sur les phénomènes statistiques qui règlent les grands réseaux ou les grands systèmes ouverts. Si l'on est en quête d'un modèle du Web, disons sa *topologie*, voilà une seconde propriété d'importance : aux côtés du constat de son *diamètre court*, le Web affiche une profonde *hiérarchisation*. L'une et l'autre propriétés constituent de puissants indicateurs d'organisation qui nous éloignent diablement aussi bien des graphes (ou des systèmes) réguliers que des graphes aléatoires qu'aiment à générer les physiciens. Sur le Web, la distribution des liens n'est donc ni uniforme ni aléatoire, mais concentrée sur un nombre réduit d'éléments qui, de surcroît, voient leur position dominante se renforcer au cours du temps. On peut donc parler d'un principe de renforcement temporel : plus le temps avance, plus augmentent les chances que les nouveaux liens pointent vers des *hubs* et/ou que les *hubs* en retour pointent les

nouveaux nœuds arrivants qu'ils sont en capacité de repérer. On comprend alors pourquoi le diamètre moyen du Web diminue malgré l'augmentation gigantesque du nombre de documents : une forme d'ordre émerge et se renforce à partir des masses de documents et semble gouverner leur évolution, de façon implacable et universelle. Lorsque l'on réfléchit aux implications d'un tel principe d'évolution temporelle, on est surpris de sa conclusion logique : à la fin de cette histoire théorique, il ne reste plus qu'*un seul* nœud central, et malgré l'expansion continue du nombre de documents. « *Winner-take-all* ». Le Web ne serait plus un réseau distribué, mais une étoile, une *web star*, unique et attirant à elle toute la connectivité. Dans *Linked*, Albert-László Barabási en caresse l'idée, se demandant qui de Google ou Microsoft gagnera la course⁵.

À ce stade, les questions se bousculent. En premier lieu, il devient évident que l'on peut lier les questions de distances moyennes (ou de diamètre) à celle de la distribution en loi de puissance. Quel que soit le nombre d'éléments d'un système et le nombre de leurs liens réciproques, l'application d'une loi de puissance permet évidemment de deviner son diamètre moyen ou bien encore la distance moyenne entre n'importe quels éléments pris deux à deux (des paires de nœuds). Ensuite, on peut esquisser une façon d'apercevoir rétrospectivement l'histoire du Web comme système ou bien, de façon prédictive, son développement futur. Né à partir de quelques pages inaugurales (le *big-bang* du système), le développement exponentiel du réseau conduit la connectivité associée aux liens hypertextes à se développer localement comme globalement, mouvement distribué et ascendant qui aboutit à l'émergence d'une forme d'ordre. Cette évolution continue se nourrit du rôle central de cette classe de nœuds ou d'éléments que sont les *hubs* qui assurent à l'ensemble son unité, sa cohésion hiérarchique (sa colonne vertébrale si l'on veut) et, peut-être aussi, sa synchronicité (puisque les distances sont courtes). Enfin, ces fameux *hubs*, piliers du système documentaire qu'est le Web, peuvent aussi constituer son talon d'Achille⁶. Un réseau, en tant que tel, n'existe plus à partir du moment où ses composants (individuellement ou par groupes) ne sont plus reliés entre eux, qu'il n'existe plus de diamètre possible faute de chemins. Le Web, comme tous les autres réseaux gouvernés par une loi de puissance, s'est révélé dès les années 1998-1999 particulièrement résistant à des attaques massives portant indistinctement sur tous les types de nœuds. D'après certains chercheurs, on peut même faire disparaître plus de 70 % des éléments d'un réseau comme le Web sans pour cela l'éliminer comme espace connecté où des chemins sont toujours possibles (mais certes plus longs). En revanche, faites disparaître les principaux *hubs* lors d'une attaque ciblée et votre

5. A.-L. Barabási, *Linked*, op. cit.

6. R. Albert, H. Jeong et A.-L. Barabási, "Error and attack tolerance of complex networks", *Nature*, 406, 2000, pp. 378-382.

réseau se disloque immanquablement en une myriade d'éléments solitaires et de petits groupes⁷. Une forme concentrée autour de quelques *hubs* fait donc à la fois la robustesse et la fragilité du réseau, et simultanément, tout comme son diamètre court malgré son nombre croissant de documents.

D'un certain point de vue, ce type de propriétés nous éloigne encore un peu plus de l'univers maîtrisé de la bibliothèque. Certes, on pourrait associer de façon contiguë les ouvrages en fonction des liens de citations, rappelant ainsi combien les méthodes d'analyse des documents web doivent à celles, plus anciennes, de la bibliométrie. Cependant, rien n'indique une telle organisation massive et globale à l'échelle de l'ensemble du système dans l'étude de la distribution des liens de citation dans les grandes bases de données de documents scientifiques, ni même à l'échelle des concepts communs (quel y serait l'équivalent des *hubs* ?), ni à celui des réseaux de collaborations scientifiques ou de copublications d'articles par plusieurs auteurs. L'étude des principes qui guident la distribution des liens hypertextes nous fait entrer dans une sorte de physique documentaire, d'autant plus nécessaire qu'il n'y a pas sur le Web de catalogue, d'index, de tables ou même de listes qui pourraient être classées et regroupées en ensembles cohérents. Il faudra peut-être abandonner définitivement le rêve du Web devenant une sorte de « grande bibliothèque universelle » (fût-elle collective).

C'est pourquoi il nous a fallu changer de modèle d'univers et regarder vers d'autres types de systèmes ou d'organisations pour établir quelques rapprochements utiles ou salutaires dans notre quête de la « géographie du Web ». Il ne fallait donc plus considérer le Web comme un système d'information, du moins au sens classique. Il fallait le regarder comme écosystème informationnel, l'*e-cosystem* tel que le qualifie Bernardo Huberman⁸. Lorsque j'ouvre l'Atelier de Cartographie en 2011, le blog s'insère presque naturellement dans un *écosystème* auquel il contribue et ce dernier, en retour, l'intègre organiquement comme élément constitutif de son architecture. Il enrichit une sorte de *biotope documentaire* déjà existant, avec ses *patterns* de *distance*, de *diamètre*, de *hiérarchies* ou de *groupes*. Nous sommes ainsi très proches du concept de biodiversité et de la façon dont la biologie et les sciences de l'environnement ont été amenées à décrire cet objet de recherche comme un ensemble d'interactions entre différentes espèces : relations de différentes natures (prédation, reproduction, coopération, symbiose, pollinisation...), à différentes échelles (micro-comme macroscopiques) et distribuées de façon hiérarchisée (comme sous la forme d'une pyramide) ou de façon latérale.

7. *Ibid.*

8. B.A. Huberman, *The Laws of the Web: Patterns in the Ecology of Information*, Cambridge (MA), MIT Press, 2001.

Directement relié aux *méga-hubs* du Web (un blog Wordpress, synchronisé avec un compte Twitter et un profil LinkedIn), il est aussi relié localement à ce que l'on peut supposer être un groupe d'acteurs de l'innovation numérique (latéralement, mes pairs que je cite aussi dans les *posts*) et en toute proximité hypertexte avec les autorités scientifiques du domaine des sciences des réseaux et de la cartographie d'information. À titre d'administrateur du blog, Wordpress me fournit d'ailleurs une sorte de *dashboard environnemental* d'où je peux estimer le bruit (ou le *buzz*) de mes posts qui agissent comme des signaux via les statistiques de fréquentation, savoir combien d'autres objets web les ont repérés ou s'y intéressent, et quel niveau de hiérarchie ils occupent (fréquentation via les moteurs de recherche, via les abonnements au blog lui-même, via Twitter...).

Image 3.6

Copie d'écran de l'interface d'administration d'un blog Wordpress.

Me voilà équipé pour insérer mes publications dans cet e-cosystème documentaire : côté pile, celui de l'interface publique, il expose des contenus qui s'empilent chronologiquement et sont reliés à d'autres contenus complémentaires qui participent à la construction régulière de la localité (ou la communauté) dans laquelle il se trouve et dont il pointe le cœur de références via le *blogroll*. Côté face, un *dashboard de monitoring* permet de mesurer statistiquement la façon dont il prend place sur le réseau en fournissant une série d'indicateurs agrégeant la dynamique

de diffusion des contenus (statistiques pour chaque post du blog) et de réception ou, plutôt, de *visibilité* (nature et intensité de l'écho généré par le système quand je publie un nouveau post).

Il nous restait présent à l'esprit que le principe d'une loi de puissance gouvernant la forme, voire l'évolution du Web, pouvait être discuté et interrogé de bien des façons. À commencer par sa « simplicité » et son « universalité » (presque suspecte d'une certaine façon) qui suppose une hiérarchisation continue de ce vaste système documentaire distribué. À l'évidence, rien ne garantit qu'il en soit toujours ainsi et peut-être ne s'agit-il là que d'une première phase d'où émerge une forme d'organisation spontanée à travers le temps, phase typique d'un système complexe en émergence. On peut aussi concevoir une sorte de révolution structurelle du système, un bouleversement rapide et généralisé des hiérarchies et des centres de gravité comme le font certains chercheurs en science des réseaux. L'équivalent, si l'on veut, de la *mutation* en biologie, de la *révolution* en histoire ou, surtout, de la *transition de phase* en sciences physiques dont les modèles permettent d'expliquer la transformation de la matière dans le passage d'un état (par exemple liquide) à un autre (solide). Un tel événement ne s'est pas (encore) réalisé pour le Web malgré son évolution technologique rapide, à supposer donc qu'il adienne et que l'on ait les moyens de l'observer. Malgré les incertitudes, il reste acquis qu'une loi de puissance régit la distribution de la connectivité sur le Web et qu'un tel principe semble gouverner bien d'autres systèmes dits « complexes ». Si l'on associe ce principe à celui des diamètres relativement courts, on comprend rapidement que l'un et l'autre constituent des *indices remarquables* de ce qui fait *système*, qui évolue par *synchronisation* de ses éléments constitutifs (qu'il s'agisse d'évolution ou de révolution). Règne ici une forme d'ordre, une *topologie* particulière. Reste à en dévoiler l'architecture, car nous sommes encore, sous de nombreux aspects, en *terra incognita*. Assurément, nous ne sommes pas dans un univers produit par le hasard (*randomness*) ou animé par des connexions incohérentes et non coordonnées, ni non plus régi par des régularités logiques qui nous sont familières comme les arborescences ou les organisations régulières (comme notre « filet de pêche » avec ses mailles sous forme de graphe régulier). Une forme d'ordre, encore à peine esquissée, mais qui ne doit rien au hasard, et à laquelle je me suis bien *naturellement* plié dès l'ouverture de mon blog, l'Atelier de Cartographie.

Chapitre 4. Agrégats

Agrégat : assemblage hétérogène de substances ou éléments qui adhèrent solidement entre eux. Syn. : agglomérat, conglomerat, Petit Robert de la langue française.

Mon bureau à l'université de technologie de Compiègne (UTC) servait certains soirs de lieu de rendez-vous pour l'équipe d'étudiants-ingénieurs avec qui je partageais mes interrogations sur une possible *carte du Web*. Un plan, voire un simple schéma aurait suffi à combler nos espoirs, enfin un modèle général plus avancé que le « nœud papillon » et qui aurait pu guider le développement de technologies originales et ouvrir de nouvelles pistes. Mes lectures sur le diamètre du Web et la découverte du principe du *small world phenomenon* m'avaient conduit à partager immédiatement avec ce groupe d'étudiants l'idée d'une exploration du Web comme un *territoire numérique*, à partir d'instruments inédits qui auraient bientôt, à n'en point douter, le retentissement équivalent à l'invention du télescope ! Nos expéditions sur ces territoires et nos premiers arpentages nous ont fourni l'occasion de nos premières mesures, de nos premiers clichés cartographiques et la découverte d'un continent encore peu connu. Le bureau ressemblait à une sorte d'atelier où de futurs ingénieurs exerçaient leurs compétences autour de ce projet insensé, de préférence la nuit pour bénéficier de toute la bande passante disponible dans le bâtiment du centre de recherche de l'UTC. La discrétion n'a pas été notre fort durant cette période de quelques mois : l'administration nous avait plusieurs fois avertis (avec raison finalement) des problèmes que nous posions en termes de « sécurité » pour les systèmes informatiques de l'université, avec ces « robots » lâchés sur le Web, et dont nous ne contrôlions pas toujours le comportement. Quelques mésaventures nous avaient cependant permis d'enrichir notre expérience et nos questions sur ces nouveaux territoires documentaires, notamment cette fois où en recensant les sites et les pages web issus de l'univers de la défense des *sans-papiers*, nos robots (les *crawlers* qui arpentent le Web de lien hypertexte en lien hypertexte) avaient dérivé systématiquement dans

l'univers des premiers grands projets *open source* de l'époque, violant ainsi les règles de protection des serveurs. Les robots ayant été lâchés depuis l'ordinateur de mon bureau, je fus donc logiquement désigné comme responsable de l'infraction.

Cela n'entama pas notre ardeur : notre atelier nocturne vivait d'une sorte de bricolage inventif permanent où chacun s'attelait spontanément à sa tâche. Les ateliers renaissants de Florence ou de Venise, où l'on apprenait aussi bien la peinture que la sculpture, la marqueterie, le plâtre et les pigments comme l'architecture, devaient présenter ce même désordre apparent : pendant que les uns développaient des instruments d'analyse des contenus des pages web (indexation *full-text*, calcul des fréquences de termes utilisés dans les documents, repérage des hiérarchies textuelles depuis les titres, les métadonnées ou les paragraphes et leurs longues chaînes de caractères...), d'autres travaillaient à la façon d'archiver sous forme de graphes les informations issues des pages (qui devenaient donc des nœuds) et des liens hypertextes tels que nos robots les découvraient, sans oublier ceux qui tentaient de trouver des modes visuels adéquats de représentation pour produire ces fameuses *cartes* dont nous rêvions. Il fallait bien forger de nouveaux instruments qui nous ont rapidement permis de sonder le Web à différents endroits comme le feraient des géologues, sans l'aide de Google ni d'un navigateur.

Une telle effervescence rappelle combien notre aventure a pu reposer en premier lieu sur la combinaison de compétences techniques et humaines, et d'une grande diversité. La réussite du projet passait par la mise en place de toute une chaîne de traitement de l'information. Celle-ci reposait sur le développement, puis l'ajustement de quatre types de briques technologiques élémentaires, car il faut pouvoir extraire l'information (un serveur de données et des *crawlers* web), indexer et archiver les données, les analyser (notamment avec des métriques de graphes qui nous semblaient l'approche la plus adaptée dans notre cas) et, enfin, une interface d'accès aux résultats dans laquelle sera intégré un espace d'exploration visuelle.

Image 4.2

Chaîne de traitement de l'information.

Extraction-indexation-métriques statistiques-spatialisation : l'alignement des quatre processus ne doit pas masquer qu'il s'agit en réalité d'un procédé *modulaire*, et très dynamique en phase de conception et d'essai. Pour chacune des quatre phases de la chaîne de traitement de l'information web, plusieurs méthodes sont possibles : il faut pouvoir imbriquer les algorithmes à partir de procédures robustes de transfert d'une brique à l'autre (*data processing*), mais aussi les aligner de façon à faire apparaître « quelque chose » à « un moment » qui ressemble à un principe d'ordre ou de redondance, sous la forme de mesures convergentes, d'un *pattern* statistique ou d'une forme visuelle. Inutile de préciser qu'il a fallu tester nombre de possibilités combinatoires pour arriver à saisir une forme d'intelligibilité dans nos premières données.

Image 4.3

Chaîne de traitement de l'information : des allers-retours sont nécessaires entre les étapes.

Cependant, il arriva un moment où, comme avec un télescope, la combinatoire diminue et s'affine (disons, les essais tous azimuts) au profit d'une série de réglages plus fins à mesure que nos premiers objets se dessinent. Enfin, une disposition des processus nous a semblé optimale quand nous avons braqué notre « télescope » sur une zone du Web curieuse et originale que pour la première fois nous avons pu observer avec une certaine précision, celle de l'Église de scientologie en ligne et de ses détracteurs.

Je ne sais toujours pas aujourd'hui à quel type de science nous participons et, hormis notre volonté d'achever le projet, aucune place n'était attribuée à notre démarche dans l'organigramme scientifique commun, ni en sciences humaines et sociales (où le projet était jugé trop « technique » et faiblement articulé théoriquement) ni, non plus, en sciences informatiques (où le projet apparaissait vite comme une sorte d'excroissance de la sociologie). À bien y regarder, il s'agissait d'une recherche technologique fondée sur l'idée que les hypothèses, les concepts, voire les théories apparaîtraient de façon ascendante depuis les expérimentations, sans prétention de départ et en assumant notre ignorance partagée. Les démarches exploratoires comme la nôtre, attendant donc que les données elles-mêmes ouvrent la voie aux hypothèses, ont le

mérite de partir du constat implicite de notre ignorance à l'égard des phénomènes étudiés, contrairement à la conception classique de la science expérimentale (à laquelle les SHS prétendent aussi), où c'est à partir de connaissances que l'on produit d'abord des hypothèses sur un objet que l'on se donne *ensuite seulement* les moyens de vérifier (ou de falsifier) expérimentalement. Je sais, j'exagère : un examen attentif des pratiques scientifiques révélerait combien les deux mouvements sont en réalité combinés dans l'activité scientifique. Cependant, je ne peux m'empêcher de penser que notre démarche inductive, comme bien d'autres à l'époque et encore aujourd'hui (concentrée dans ce nouveau domaine appelé *data sciences*), font symptôme en affirmant le rôle prépondérant des instruments (et donc des ingénieurs, au moins dans un premier temps), en pratiquant une sorte d'ascèse devant les données de façon à « oublier » les connaissances acquises antérieures. Les prouesses technologiques exploratoires sont les produits d'une forme d'ignorance ou de questionnement de départ, et non des garanties théoriques issues des mondes de l'érudition, de la répétition, des certitudes idéologiques. Dans certains contextes, ces types de démarches exploratoires se multiplient comme les symptômes d'un savoir en crise, de modèles théoriques dépassés, d'anciennes écoles de pensée s'affaiblissant au contact d'une réalité contemporaine devenue alors insaisissable, incompréhensible, complexe, imprévisible. Nos experts en économie ont-ils su prédire les crises financières des années 2006 à 2010 ? Les sociologues savent-ils envisager autrement le monde social qu'à travers la dichotomie *micro-macro*, autrement que par des méthodes qualitatives ou quantitatives ? Les technologies classiques d'indexation de contenus des documents ou les instruments arborescents de gestion de l'information peuvent-ils rendre compte de la forme du Web, le plus grand réseau de documents jamais réalisé ? Toutes ces masses de données associées au *big data*, hétérogènes et multidimensionnelles, peuvent-elles être traitées avec les statistiques habituelles et leurs modèles linéaires ?

Ce type de questions fait désormais partie du paysage intellectuel en *data sciences*, dont l'une des bases méthodologiques est d'être *data driven*, conduit par les données et l'expérimentation. À l'époque de nos premières expéditions sur le Web, personne ne pouvait imaginer que l'on puisse faire métier d'une telle activité exploratoire ! La première vue de la topologie où nous entrions avait à l'écran un aspect fascinant et prometteur :

Image 4.4

Exploration du Web de l'Église de scientologie. F. Ghitalla, E. Diemert, C. Maussang et F. Pfaender, "TARENTE: An experimental tool for extracting and exploring web aggregates", dans *Proceedings. International Conference on Information and Communication Technologies: From Theory to Applications*, IEEE, 2004, pp. 627-628.

Dans cette vue produite par Eustache Diemert au cours de l'année 2003, chaque point représente une page web du *set* primaire de données extraites avec un *crawler*. L'ensemble est projeté sur une surface circulaire en 3D. Après un processus d'analyse et de sélection, seules les pages qui entraient dans la thématique choisie (la scientologie, ses adeptes et ses détracteurs) figurent ici en noir ainsi que les liens hypertextes qui les relient. Dans la partie haute de la vue, figurent par exemple des documents émanant de différents acteurs individuels ou collectifs (associations, sites web de députés ou de sénateurs, anciens adeptes de la scientologie...) reliés entre eux et qui partagent des fréquences de termes communes et assez élevées comme « secte », « dérives », « Tom Cruise » ou « dangers ». En bas, ramassé sur lui-même, figure le *cluster* des adeptes de l'Église, très nombreux et qui renvoient tous vers les sites officiels (américains) de la secte. Ce caractère très concentré de la distribution des liens internes à la communauté des « croyants » s'est révélé à l'analyse très hiérarchisé : il y a presque quinze ans, à l'époque du Web *statique* 1.0, l'Église fournissait à ses adeptes-internautes des modèles de sites préformatés où les liens externes étaient déjà présents (pointant notamment directement sur les références officielles de l'Église à Ron Hubbard, son fondateur). On peut supposer que les responsables de la scientologie et de sa communication étaient déjà pleinement conscients de la façon de « se donner » au Web sous la forme d'une organisation se dotant d'une topologie particulière, presque organique. À l'examen, seul un *gateway* (en rouge) permet de relier les documents de la communauté au reste du Web qui comporte pourtant plusieurs centaines de millions de pages à l'époque. Ce *gateway* vers la scientologie

en ligne n'est pas anodin : il était constitué d'un ensemble de documents juridiques issus des jugements de différents tribunaux (en France et aux États-Unis notamment) en faveur de la scientologie (donc dans des pages web pointées par les sites de l'Église), mais aussi en sa défaveur (donc d'autres pages pointées par ses détracteurs). Notre *gateway* de documents juridiques est donc seulement alimenté par des liens entrants pointant soit vers certaines de ses pages faisant mention de jugements favorables à l'Église de scientologie soit, pour les autres liens hypertextes, vers des pages qui lui étaient défavorables. De part et d'autre de cette ligne de démarcation, les documents parlent ou d'« Église » ou de « secte ».

La description détaillée de ce premier cliché cartographique ne relève pas simplement de l'anecdote ; il s'agit d'un *premier cas* dont l'essentiel des propriétés se sont trouvées confirmées dans les mois qui suivirent à partir d'une foule d'études diverses et d'observations. Et partout la même *structure granulaire*. Il n'est guère possible aujourd'hui de recenser toutes nos explorations, mais ce type de sondage de la structure documentaire du Web s'est rapidement étendu depuis les premières cartographies dédiées aux communautés compétitives, impliquées par exemple dans des controverses scientifiques et techniques (là où les adversaires en ligne ont tendance à former des *clusters* particulièrement nets et précis), jusqu'aux phénomènes globaux plus généraux comme l'expression politique, les organisations de la société civile, l'avortement (*pro-life*, *pro-choice*), les ressources liées au domaine de la CSTI (Culture scientifique, technique et industrielle), les adeptes des jeux en ligne et leurs plateformes (les ancêtres de Steam). Mais nous avons aussi sondé des domaines dont les portes d'entrée étaient accessibles sur le Web comme Wikipedia, les sites web de certaines universités. Partout se vérifie un principe d'agrégation de pages web qui échangent entre elles un nombre important de liens hypertextes et qui partagent la même thématique. Le principe semble aller de soi, et l'observation de l'activité des internautes le confirme : chacun se rassemble et partage avec toutes sortes de pairs et de proches des informations communes ou complémentaires. Cela fait les délices de ces nouvelles Computational Social Sciences (CSS) en vogue actuellement. Mais au début, on n'a guère envisagé, en revanche, les implications de ce principe fort pour la géographie documentaire du Web et des instruments qui lui sont associés.

À commencer par le principe de production d'*agrégats de documents* sur le Web¹. « Agrégat : assemblage hétérogène de substances ou éléments qui adhèrent solidement entre eux. Syn. : agglomérat, conglomérat », comme le définit le Petit Robert de la

1. À ma connaissance, la première apparition du principe d'agrégation dans les données a été décrit en 1991 par R. A. Botafogo et B. Shneiderman, *Identifying Aggregates in Hypertext Structures*, dans HYPERTEXT '91: Proceedings of the third annual ACM conference on Hypertext, 1991, pp. 63-74, <https://doi.org/10.1145/122974.122981>.

langue française. Et cette définition correspond exactement aux phénomènes que nous avons étudiés de nombreuses fois où la *proximité hypertexte* induite par la distribution des liens se double – ou se synchronise – avec une *similarité sémantique* à travers le partage de mots-clés communs. C'est comme si le Web était constitué, entre le nuage et les abîmes, en dessous de *Bing*, *Yahoo*, *Google* ou *Facebook*, mais au-dessus du *deep web*, d'une couche médiane constituée par une multitude d'agrégats de documents. Comme une structure granulaire, le ciment de l'architecture documentaire distribuée. Et c'est cette couche médiane que nous avons sondée de nombreuses fois, en essayant d'y naviguer de façon « latérale » comme on le ferait pour estimer l'étendue d'une nappe phréatique.

Image 4.5

Modèle du Web en couches.

Un schéma cohérent du Web apparaissait enfin, compatible avec nos premières intuitions : entre une couche « haute » gouvernée par les grands distributeurs de liens comme les moteurs de recherche et les réseaux sociaux et une couche « basse » documentaire, la couche médiane et sa structure granulaire où les deux premières « couches » se répondent et se traduisent mutuellement à travers la dynamique sociale de l'agrégation. En haut, une puissante machinerie générant l'essentiel de connectivité du système ; en bas, la logique de production documentaire des contenus et, au milieu, la dynamique sociale de l'agrégation en *clusters* identifiables. Cette première version du schéma du « modèle du Web en couches » était compatible aussi bien avec les données expérimentales que nous accumulions qu'avec les premières hypothèses diffusées aux États-Unis depuis les années 1997-1998. En effet, dès les années 2000, des chercheurs d'IBM regroupés autour du projet CLEVER au Almaden

Research Center² avaient déjà isolé le principe de l'agrégation sous la forme de l'étude des corrélations contenu-structure (hyperliée). Leur objectif n'était pas de produire une géographie ou un modèle topologique du Web, mais seulement d'optimiser les méthodes de traitement de l'information pour les moteurs de recherche. Résumée à grands traits, il s'agissait de repérer rapidement des pages pertinentes (par exemple par rapport à une requête) en analysant à la fois son contenu et la façon dont elle était reliée au reste du Web. On a donc rajouté aux tables d'indexation des contenus des documents web des scores de connectivité pour les hiérarchiser, avec une attention prononcée chez les fondateurs de Google pour les pages « faisant autorité » dans un agrégat, autrement dit ayant de nombreux liens entrants ou de chemins y menant. Après tout, il n'y avait qu'à exploiter cet immense système spontané de sélection ou de vote collectif auquel j'ai participé en ouvrant mon blog l'Atelier de Cartographie, avec son *blogroll* tourné vers mes sites de référence. D'une certaine façon, pointer c'est voter ou ratifier. Mais les techniques de traitement de l'information web documentaire peuvent devenir plus complexes en alternant les phases de sélection des pages à partir de la distribution de leurs liens hypertextes réciproques d'un côté, de l'analyse de leurs contenus de l'autre. On peut même comparer les deux approches en traitant de façon parallèle *le même corpus* de documents identifiés grâce aux *crawlers* : a) en commençant par l'analyse des liens, puis en finissant par l'analyse des contenus, b) en commençant par l'analyse des contenus, puis en finissant par l'analyse des liens. C'est exactement ce que nous avons tenté avec notre corpus de pages extraites de l'univers de la scientologie.³

Image 4.6

En 2004, Camille Maussang, Eustache Diemert et Fabien Pfaender ont testé deux chaînes de traitement d'un corpus de pages web et comparé les résultats finaux dans l'identification d'agrégats de documents web. Les deux processus parallèles (A des liens aux contenus, B des contenus à la structure hyperliée) sont apparus convergents, tant dans la définition du cœur (*core*) de l'agrégat que dans celle de ses frontières (*borders*).

2. <http://www.research.ibm.com/labs/almaiden/index.shtml>.

3. F. Ghitalla, E. Diemert, C. Maussang et F. Pfaender, "TARENTE: An experimental tool for extracting and exploring web aggregates", dans *Proceedings. International Conference on Information and Communication Technologies: From Theory to Applications*, IEEE, 2004, pp. 627-628.

Les deux voies ont conduit à identifier la même organisation dans les pages ! Avec les bons réglages, il devenait donc théoriquement possible de balayer la couche médiane des agrégats pour identifier toutes les thématiques autour desquelles se regroupent les internautes. L'idée de pouvoir en posséder le catalogue complet esquissait donc une piste pour réduire le problème des masses de données web, leur caractère hétérogène et dynamique dans le temps. Une façon de le borner, de retrouver la plénitude de l'exhaustivité, la maîtrise des frontières de contenus, peut-être même l'arbre de leur hiérarchie... en somme, l'espoir d'un plan détaillé de la distribution à grande échelle de *tous* les contenus documentaires du Web.

Ce sont les chercheurs et les ingénieurs d'IBM qui ont ainsi pour la première fois décrit le Web comme une architecture spatiale, certes inédite, mais potentiellement réductible à un modèle, une topologie. En son principe, la théorie des agrégats de documents web suppose que ce qui se ressemble thématiquement se rassemble hypertextuellement, autrement dit que fusionnent *similarité* et *proximité*. C'est ce processus d'agrégation qui expliquerait cette fameuse force de gravitation déjà rencontrée : les documents de cet univers viendront d'eux-mêmes s'agréger à leurs agrégats thématiques, renforçant ainsi ceux déjà présents, et surtout les fameux *hubs*. C'est ce que les chercheurs californiens de Palo Alto ont désigné par *topical localities* et leur ingénierie des données un procédé de *topical distillation*⁴. Avec l'étude sur la scientologie, nous avons donc identifié deux localités thématiques, et proches l'une de l'autre. Tout commençait à faire sens, y compris les petits détails comme notre mésaventure avec nos premiers *crawlers* qui avaient quitté la localité des « sans-papiers » pour migrer dans un agrégat voisin, celui des projets informatiques *open source*. Les rapprochements thématiques que nous propose le Web (son système de distribution des contenus dans la structure hyperliée) peuvent aussi réserver quelques surprises, parfois déroutantes, comme ces liens qui arrivaient vers les meilleurs *hubs* des détracteurs de la scientologie et qui provenaient... de sites catholiques et d'organisations chrétiennes.

Parmi tous ces travaux sur les principes de corrélation contenu-structure, ceux de Jon Kleinberg ont selon moi un statut particulier. Au cours de l'année 1997, il conçoit un algorithme particulier qui permet d'isoler la structure interne d'une localité thématique sur le Web⁵, autrement dit un agrégat. HITS – Hypertext Induced Topic Search – restera une contribution majeure du domaine du *web mining* en ce qu'il concentre de façon synthétique et habile un ensemble de problématiques méthodologiques

4. K. Bharat et M.R. Henzinger, "Improved Algorithms for Topic Distillation in a Hyperlinked Environment", dans *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, p. 104.

5. J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", <http://www.cs.cornell.edu/home/kleinber/auth.pdf>.

complexes en une seule formule (M.E.J. Newman dira de cet algorithme qu'il représente une *elegant construction*⁶). Il faut imaginer HITS comme une sorte de microscope appliqué à ces agrégats de documents web où les documents partagent entre eux une forte densité de liens et, simultanément, certaines similarités de contenu. Kleinberg met au point un calcul qui fonctionne comme une sorte de balance entre liens pointant vers une page web et liens sortant de la même page. Si l'on prend un corpus de pages web traitant d'un même sujet (on peut prendre une liste de résultats sur Google générée à partir d'une requête), HITS permet d'appliquer à chacune des pages du corpus un score d'*authority* et un score de *hub*. Si l'on regarde les scores obtenus, on s'aperçoit de deux phénomènes majeurs. En premier lieu, comme le laissait supposer le principe générique de la loi de puissance, la distribution des scores en fonction des liens entrants et sortants est très hiérarchisée. On peut donc isoler, dans un agrégat de documents, un nombre limité de nœuds au cœur de la distribution de la connectivité hypertexte. Le second principe consiste à examiner en détail ce cœur, par exemple sous la forme de deux listes – l'une faisant figurer les pages en fonction des meilleurs scores d'*authority* et l'autre en fonction des meilleurs scores de *hub*. Il s'agit de prendre en compte les liens entrants pour définir des *authorities* et les liens sortants pour définir des *hubs* en prêtant attention à la façon dont les deux se répondent et se renforcent. Les deux classements des pages web d'un agrégat en fonction du calcul de leurs scores de *hub* et de leurs scores d'*authority* apparaissent différents, parfois très nettement. Le cœur d'un agrégat est donc constitué de deux types distincts d'éléments qui se renforcent mutuellement: les *hubs* qui pointent sur tous les documents de l'agrégat, mais se concentrent sur les *authorities* du domaine, *authorities* que l'on distingue d'autant plus qu'elles sont pointées par les meilleurs *hubs*. On connaît la suite extraordinaire de l'application de l'identification des *authorities* à l'échelle du Web avec Google, et son PageRank.

6. M.E.J. Newman, *Networks: An Introduction*, Oxford, Oxford University Press, 2010.

Image 4.7

Les *hubs* et *autorités* détectées par l'algorithme HITS d'après le principe de renforcement mutuel.

HITS est plus qu'un algorithme: il contient en germes une vision du Web, un modèle organique qui ouvre sur des questions fondamentales de topologie du réseau, mais aussi de scénarios d'évolution puisque, comme le pense Jon Kleinberg, la mutualisation des liens entre *hubs* et *authorities* se renforce au cours du temps. Encore une fois, on retrouve le principe de l'augmentation continue des forces locales de gravité malgré l'extension sans fin de l'univers web. Dans son article fondateur, l'auteur inscrit sa démarche dans la lignée de la bibliométrie et des mesures de citations d'auteurs à auteurs dans le domaine des publications scientifiques. Il inscrit aussi sa contribution comme un élément méthodologique nouveau dans l'étude des réseaux sociaux. Mais les contributions de Jon Kleinberg ont trouvé une résonance particulière dans les domaines du *data mining* et des *web sciences*, en raison de l'efficacité de l'algorithme HITS à réduire les masses de données web à un processus précis d'agrégation et qui permettait, enfin, de comprendre la dynamique de l'architecture documentaire du réseau. Si l'on développe les principes esquissés par HITS, on peut admettre que le Web peut être étudié sous la forme d'un graphe dont on pourrait identifier les zones d'agrégation hypertexte pour recenser l'ensemble des « sujets » (*topics*) qui y figurent, puis les analyser pour en extraire un cœur de ressources pertinentes (*authorities*) ou celles qui constituent le meilleur aiguillage (*hubs*) si l'on veut découvrir le domaine.

Image 4.8

Un « cœur » de ressources composé de pages ayant, d'un côté, les meilleurs scores de *hub* et, de l'autre, les meilleurs scores d'*authority*. Les éléments de ce cœur (bipartite) voient leurs relations mutuelles se renforcer : a) au cours du temps, b) avec l'arrivée de nouveaux éléments dans le système, c) sous l'effet de calculs associés à HITS (ou à bon nombre d'algorithmes du même type) qui peuvent pondérer le poids des liens en fonction du score de *hub* et/ou d'*authority*. La distribution hiérarchique des liens sur le Web (parfois en « loi de puissance ») permet d'isoler les ressources centrales dans un set de données. Le procédé permet : a) d'inférer l'existence de communautés thématiques de documents à partir de l'étude des liens, b) de classer des ressources documentaires identifiées au préalable pour leur proximité de contenu. À plus grande échelle, le principe de l'agrégation locale (connectivité) de documents relativement similaires (focus sémantique) conduit à concevoir le Web documentaire comme un *espace granulaire*.

Le succès de la démarche de Jon Kleinberg tient au fait que HITS semble épouser au mieux cette dimension spécifique du Web où se croisent contenus et distribution des connexions (les deux « manettes » dont s'emparent les internautes pour publier). Le principe d'une *granularité documentaire* qu'il permet de dévoiler dans les *web data* confirme (et non *produit*) l'effet de convergence entre distribution des liens hypertextes et analyse des contenus. La granularité médiane est donc un élément central de la topologie du Web. La démarche de Kleinberg s'inscrit dans une période clé des technologies de l'information et des réseaux où émerge à partir des années 1996-1997 une sorte de géographie des agrégats de documents web qui nourrit les réflexions théoriques sur la topologie du système comme les méthodes d'observation, voire de cartographie, des informations sur le Web. Au-delà de HITS, passé à la postérité, c'est tout un ensemble de travaux et de découvertes qui s'inscrit dans ce mouvement tel que l'illustrent les contributions des collègues de Jon Kleinberg du projet CLEVER comme David Gibson⁷, Soumen Chakrabarti (l'auteur de *Mining the Web: Discovering Knowledge from Hypertext Data* en 2002⁸), de Byron Dom et de Ravi Kumar, Prabhakar Raghavan et Sridhar Rajagopalan⁹ (qui publieront plus tard en 2000 avec Andrei Broder la fameuse étude sur le modèle du Web en « nœud papillon »). On doit à ce courant de recherche une série de concepts tout à fait originaux dédiés à la description du comportement des technologies (comme les *crawlers*) face à une structure documentaire de type granulaire *topical localities* (ou localités thématiques) ou les effets de *topical drifting* quand on passe d'un agrégat à un autre voisin (souvent sans le vouloir).

Cette vision très géographique, ou plutôt *topologique* du Web inaugure une série d'avancées pour les techniques de recherche d'information, qu'elles soient manuelles ou automatiques. En termes de méthodologie de veille sur le Web, l'identification rapide des bons *hubs* permet de trouver rapidement les bonnes *authorities* d'un domaine. Mais seulement si le thème étudié fait bien l'objet d'un processus d'agrégation, ce qui n'est pas toujours le cas : la façon dont s'organisent les agrégats sur le Web peut correspondre, ou non, à l'idée que l'on s'en fait. Une idée serait de balayer tout le Web « médian » pour recenser les thématiques donnant lieu à un tel processus d'agrégation pour faire « remonter » une classification que l'on peut alors qualifier d'*ascendante* (contrairement aux classifications réalisées a priori ou héritées de l'histoire comme dans les bibliothèques). C'était l'une des idées majeures associées

7. D. Gibson, J. Kleinberg et P. Raghavan, "Inferring Web Communities from Link Topology", dans *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 1998.

8. S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*, Burlington, Morgan Kaufmann, 2003.

9. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan et S. Rajagopalan, "Automatic resource list compilation by analyzing hyperlink structure and associated text", dans *Proceedings of the 7th International World Wide Web Conference*, 1998.

aux contributions méthodologiques de Jon Kleinberg : au-delà des informations typiquement contenues dans un agrégat, HITS ouvre sur la perspective de la découverte des *web communities*, autrement dit de tous les acteurs qui contribuent (volontairement ou involontairement) à alimenter de savoirs acquis ou de connaissances partagées. Difficilement admis par les documentalistes ou les bibliothécaires de l'époque, le principe de l'émergence de classifications ascendantes règle aujourd'hui bon nombre de technologies web, notamment les réseaux sociaux et la géographie de leurs communautés actives.

Les conclusions auxquelles nous ont conduit nos expérimentations comme la lecture des travaux des chercheurs californiens nous ont amené au final à dissocier le Web en trois couches fonctionnelles.

Image 4.9

Schéma du « modèle du Web en couches » réalisé en 2005 par Mathieu Jacomy. En haut, les trois couches constitutives de ce que l'on peut appeler la topologie du Web, depuis les « couches basses » (celles des systèmes d'information ou des bases documentaires) jusqu'à la couche haute, celle dans laquelle on placerait aujourd'hui Twitter ou Facebook comme sources très dynamiques de production de la connectivité. Entre les deux, la couche centrale, celles des agrégats de documents web où s'exercent des principes de corrélation forte entre la couche basse (celles des contenus informationnels) et la couche haute (celle de la structure des liens constituée par les liens hypertextes, mais aussi par les *like* et tous les systèmes de signalisation).

Ce schéma final constitue encore aujourd'hui une synthèse graphique qui concentre bon nombre de concepts scientifiques et de lois statistiques que nous avons rencontrés. Il y avait là nos trois types de « couches », surtout celle des agrégats de documents web envisagés sous différents angles méthodologiques. En particulier, en bas du schéma principal, figurent les différents types de graphes que l'on rencontrait en phase d'exploration et qui indiquaient différents types de situation problématiques : manque de données, données peu structurées, structures « en étoile » typiques de la couche haute... Ce schéma incarne pour moi une forme de rupture, en termes de méthodologie de traitement des *web data* comme des instruments dédiés à leur exploration. En particulier, il nous a contraint désormais à envisager notre objet, le système web, comme une série de couches (ou de *layers*) qui, une fois superposées, imprime au Web son organisation à titre de réseau documentaire. Nous ne faisons là que rejoindre un courant diffus, mais massif dans plusieurs champs scientifiques qui eux aussi, quelques années avant les *web sciences*, ont opéré une mutation dans la façon d'envisager leurs objets respectifs : la biologie et l'écologie naissante qui considéraient la biodiversité sous la forme de réseaux d'interactions entre espèces naturelles, la génétique et la biologie qui fragmentent le vivant depuis les gènes jusqu'aux populations sous la forme là aussi de couches reliées entre elles par des mécanismes complexes d'interdépendance.

Cette propension à concevoir différents champs de questionnement scientifique comme des systèmes (voire des *systèmes de systèmes*) illustre, me semble-t-il, les effets produits par le développement des technologies *data*, soit parce que l'on cherche à discerner dans les masses de données des « plans fonctionnels » ou des dimensions segmentées, soit, inversement, parce que le paysage prend de l'ampleur puisqu'on a fusionné différents types de *sets* de données que personne n'avait croisés jusque-là. C'est l'un des mouvements majeurs qui ouvre sur de nouvelles façons d'envisager les « systèmes complexes », nous rappelant ainsi combien la maîtrise des masses ou des quantités peut conduire à un *changement qualitatif* de l'objet de science suivant les échelles considérées. Si ce mouvement théorique dans l'étude des *web data* était prometteur, il eut aussi des conséquences sur notre façon de concevoir nos instruments. En un sens, il nous fallait donc changer nos filtres analytiques, et peut-être notre appareil cartographique lui-même...

Chapitre 5. Méthode et instruments

« Some questions have resisted answers for decades. Yet something has changed in the past few years. The driving force behind this change can be condensed into a single word: *data* [...] As scientists sift through these mountains of data, we are witnessing an increasing awareness that if we are to tackle complexity, the tools to do so are being born right now, in front of our eyes. The field that benefited most from this data windfall is often called network theory, and it is fundamentally reshaping our approach to complexity », Albert-László Barabási, “Scale-Free Networks: A Decade and Beyond”, 2009.

Lorsque Florian Arend vint dans mon bureau présenter son prototype GeneExplorer, j'étais impatient de découvrir un nouvel univers de données, celui de la génétique. Quelques mois auparavant, nous avions décidé de vérifier *ailleurs* que sur le Web le principe de l'agrégation, à grande échelle et sur des données non documentaires. L'idée n'était pas de devenir nous-mêmes biologistes ou généticiens, mais seulement de capitaliser de nouvelles hypothèses de travail, des pistes méthodologiques à tester, des améliorations techniques à apporter à nos instruments d'exploration du Web. Enfin, tout ce qui concourt à faire de la science une pratique d'atelier ou un exercice de bricolage inventif et innovant. Dans une telle perspective, emprunter des données d'un domaine scientifique voisin est chose courante à des fins expérimentales et peut ouvrir de nouvelles pistes de recherche ou de nouvelles approches. Nous nous étions tournés vers le Dana-Farber Institute à New York, spécialisé dans la recherche sur le cancer, où Marc Vidal conduit un projet fascinant : l'*interactome*¹. En son principe, l'*interactome* vise à étudier les interactions entre différentes dimensions du vivant sous forme de graphes pour essayer de comprendre les multiples mécanismes

1. http://interactome.dfci.harvard.edu/V_hostome/.

de transformation qui nous font passer du génome – c'est-à-dire un code composé des quatre bases A, C, G et U – aux êtres biologiques complexes que nous sommes et aux pathologies auxquelles nous sommes confrontés. Les données qui nous ont été prêtées proviennent de la « première » couche d'interactions complexes, celles des relations entre les gènes et les protéines qu'ils codent. Il est donc possible de produire une sorte de carte réunissant à la fois les gènes et les protéines reliés entre eux dès que les premiers contribuent à coder les secondes. Dans le cas de figure où les données relationnelles concernent deux types d'information (graphe bipartite), il est possible de construire trois types de cartes. La première est la plus simple, la carte gènes-protéines. Cependant, à partir du même jeu de données, deux autres types de cartes sont possibles : la carte gènes-gènes (tels qu'ils sont assemblés pour coder les mêmes protéines) et la carte protéines-protéines (telles qu'elles sont codées par les mêmes gènes). Notre idée était avant tout de visualiser la façon dont les gènes, réunis en assemblées, contribuent à coder les *mêmes* protéines. Nous avons donc construit une carte gènes-gènes. Nous nous demandions si les gènes se regroupent selon les mêmes principes que les agrégats de documents web. Ont-ils eux aussi élu des *hubs* ? Peut-on les clusteriser et ainsi les grouper selon des ensembles fonctionnels ? Un diamètre calculable est-il possible, du moins si une structure globale émerge ?

Image 5.2

Projet Gene Explorer.

Lorsqu'apparurent les premières vues des 4 403 gènes reliés entre eux par plus de 14 000 liens, nous avons été surpris de voir un univers *familiier*. Tout se passait comme si nos instruments et nos méthodes avaient été profilés pour l'exploration des données en biologie, alors que nous poursuivions le projet de dessiner les contours d'un système d'information (enfin, si on considère le Web comme tel). Les premières mesures ont rapidement confirmé notre impression : parmi les gènes,

les rôles étaient distribués selon le même principe général que pour nos pages web, avec des *hubs* (des gènes qui participent au codage de nombreuses protéines), des assemblées locales (des *clusters*), une distribution (presque) en loi de puissance des liens entre les gènes. Quelques groupes détachés de l'ensemble, ne dépassant pas la dizaine de gènes, flottaient en périphérie participant (probablement) directement à une structure encore plus complexe d'un niveau supérieur de *layers biologiques* comme une molécule ou une cellule. Nos connaissances étant trop limitées en génétique et en biochimie, nous n'avons guère poursuivi cette exploration. Le développement de GeneExplorer s'est alors interrompu et, comme instrument, il fait aujourd'hui partie d'une collection, une sorte de boîte à outils collective dans laquelle on peut piocher pour cartographier aussi bien les brevets que les notices de publications scientifiques ou encore des réseaux d'entreprises en territoires. Théoriquement, en revanche, cette modeste expérience nous a permis d'envisager nos données web sous un jour nouveau, de façon plus synthétique, plus organique.

Ce que la biologie réunissait, à savoir des couches morphologiquement distinctes dans la fabrique du vivant, les *web sciences* avaient pour objectif de les discerner dans leurs propres jeux de données. Il m'apparaît nettement aujourd'hui que la biologie et les *web sciences* ont emprunté des chemins inverses et complémentaires, pour arriver au même principe de partition des *data* en *layers* distincts ou en couches fonctionnelles formant un vaste écosystème d'interactions. La première a connu et vit encore de l'interconnexion de bases différentes de connaissance pour forger des *patterns* de niveau général ; les secondes ont été confrontées *immédiatement* à la question des masses et de leur dynamique temporelle (notamment les données indexées dans les moteurs de recherche). Au final, le paysage qui se dessine aujourd'hui en matière de science et de *data* semble similaire dans ses principes, autant pour la biologie que pour les *web sciences*. Évidemment, on ne peut tirer directement des enseignements d'une science pour en alimenter une autre, les spécificités des objets ou des domaines reprenant rapidement le dessus. Cependant, les rapprochements semblent nombreux et inspirants pour une démarche exploratoire comme la nôtre. J'aimais ainsi considérer les agrégats de documents web comme des sortes de protéines dans un réseau où elles s'associent pour se grouper en structures biochimiques complexes et dans des configurations spatiales spécifiques. Si la comparaison trouvait rapidement ses limites, il n'empêche qu'il s'agissait de pointer là un phénomène propre à de nombreuses sciences : si les gènes codent les protéines (les deux couches sont donc reliées), celles-ci appartiennent à l'univers biochimique gouverné par ses propres lois (assemblages complexes de protéines, fonctions de régulation, de mobilité, rôle membranaire...), tout comme les gènes eux-mêmes obéissent aussi à leurs propres lois. En quelque sorte, en passant de l'une à l'autre couche, nous changeons *scientifiquement d'objet*, tout en étant dans un *même* système. La

description et la compréhension des agrégats de documents web sont aujourd'hui remarquablement moins riches que celles des protéines : pourtant, il y aurait bien des phénomènes à étudier dans cette couche médiane du Web, les principes de corrélation entre les contenus et la structure pouvant être nombreux et dynamiques. Ce principe d'hybridation imprime à l'histoire du Web et aux façons de se le représenter l'alternance d'éclairages différents : le Web a d'abord été défini comme un système documentaire ou un système universel d'information (on a même annoncé l'avènement d'un Web sémantique universel, sa couche basse) avant le grand tournant des années 2000 et l'émergence du Web 2.0, celui des réseaux sociaux et contributifs fondé techniquement sur la distribution de liens en fonction des proximités sociales. D'autres âges viendront s'y ajouter, peut-être celui du *cloud computing* (avec ses *big data*, ses algorithmes et son interopérabilité) ou alors celui des objets connectés, le Web 3.0. Dans tous les cas, il y a fort à parier qu'y alterneront des conceptions du Web (ou de toute forme de réseau) orientées soit vers la connexion, soit vers l'information. Et entre les deux, toujours et partout le même principe d'agrégation.

Il me semble que l'on n'a guère exploité toutes les implications théoriques des travaux de Jon Kleinberg. Au-delà du domaine restreint de l'histoire des algorithmes en *web mining*, le principe des agrégats de documents web permet pourtant maintenant d'intégrer différentes propriétés du Web en un tableau cohérent. Ouvert et décentralisé, il apparaît comme un écosystème documentaire dont le diamètre diminue au cours du temps malgré sa « masse » en expansion (l'équivalent du nombre de documents potentiellement accessibles). À l'examen, son architecture apparaît comme constituée de niveaux ou de couches que leur nature statistique (en termes de liens comme de contenus) différencie nettement : une couche « haute » constituée des éléments essentiels et proéminents, les *hubs* qui génèrent et qui attirent l'essentiel de la connectivité du système lui assurant des distances courtes, une « couche » basse – presque primaire – celles de l'information au sens documentaire, avec ses milliards de lignes de caractères, les textes accompagnés de leurs images et de leurs objets multimédias aux multiples contours. L'essentiel du système, son cœur, réside dans une couche « médiane » où s'hybrident sous forme d'agrégats reconnaissables proximité hypertexte et similarité sémantique. La couche médiane n'est pas une strate de plus parmi d'autres, mais une zone de traduction dynamique et réversible entre les liens (qui assurent la navigation) et les contenus (qui portent l'information). Ce sont les contributions spontanées des internautes qui construisent cette couche médiane des localités thématiques définies, théoriquement dénombrables.

Image 5-3

La **couche haute** attire l'essentiel de la connectivité du système, y compris depuis la couche basse.

La **couche basse** contient l'information au sens documentaire. On y accède depuis la couche intermédiaire.

La **couche médiane** ou couche d'agrégation est une zone de traduction dynamique où s'hybrident similarités sémantiques (thèmes) et proximités hypertextuelles (liens).

Chaque couche du Web a des propriétés distinctes.

Une vue rétrospective générale sur le système permettrait sûrement d'en apercevoir l'évolution comme une forme de développement continu. Pourtant, il me semble que cette couche hybride est née dès l'apparition de la première page du Web, celle de Tim Berners-Lee. Ce *big bang* du Web a ouvert un espace où deux ordres se sont trouvés mis en relation pour la première fois : celui d'une logique documentaire déjà ancienne (avec ses mises en forme sémiotiques, ses langues, ses supports de connaissances, ses structurations arborescentes) et celui de la distribution ouverte de la connectivité hypertexte. Et le langage HTML, comme espace d'écriture, semble jouer dans cet espace d'hybridation le rôle d'une clef de voûte où s'écrivent autant les lignes du texte que les ancres des liens hypertextes ou les objets *embedded* en JavaScript qui rendent nos pages web si interactives aujourd'hui. On peut ainsi voir le Web comme un immense système de traduction, ou plutôt de *transition*, de la couche haute à la couche basse et *vice versa*, des contenus en connexions et des connexions en contenus. L'approche *data* qu'ont développée Jon Kleinberg, Duncan Watts ou Albert-László Barabási conduit vers une conception dynamique de la complexité du Web, et peut-être de tout système dit « complexe ». Dans ce type de conception, il faut imaginer que la complexité se loge partout là où deux ordres ou deux types de rationalité entrent en contact, des contenus aux liens hypertextes, du génotype au phénotype, du sujet social aux formes de société. Entre les deux, le contact s'établit sur une série de couches qui chacune transforme les données et semble souvent obéir à une série de lois qui lui sont propres, sans pour autant abandonner ses interconnexions avec les autres couches.

Cette conception du Web en couches distinctes oblige nécessairement à revoir quelques-uns des instruments théoriques attachés à la description du Web. C'est le cas, en particulier, de la fameuse *loi de puissance* qui fait apparaître dans la distribution des liens hypertextes une profonde hiérarchisation entre les nœuds d'un système. Globalement, elle se vérifie dans tous les jeux de données web dès que l'on passe une certaine quantité et on doit la considérer, tout comme la mesure du diamètre et les calculs de chemins, comme la signature statistique d'un univers régi par des lois et une architecture donnée, d'un certain type d'ordre sous-jacent. Mais ce n'est qu'une *signature*, l'indice d'une organisation ou d'une topologie qui attend toujours d'être expliquée, cartographiée. Rien de plus que le signe que l'exploration en détail du système peut commencer.

Le concept de loi de puissance, notamment sa représentation visuelle sous forme d'une courbe typique, jouit aujourd'hui d'une certaine notoriété avec sa redécouverte sous l'effet de l'émergence de la science des réseaux, de la puissance de calcul de nos machines et de la disponibilité de nombreux jeux de données accessibles sur le Web. Si, comme le pense Albert-László Barabási, plus de 80 % des systèmes dits « complexes » (naturels, techniques, sociaux, économiques...) obéissent globalement à un principe de hiérarchisation spontanée de leurs composants ou de leurs éléments, je dois avouer qu'il y a de quoi être étonné par ce principe d'ordre qui semble s'appliquer à tout l'univers, de la plus petite à la plus grande des échelles. Le concept de loi de puissance est d'autant plus fascinant que chacun peut aujourd'hui en vérifier la validité sur tous les domaines documentés par les technologies de l'information, depuis les degrés d'occurrence des *keywords* dans les publications scientifiques jusqu'aux mouvements des capitaux dans une économie mondialisée, sans oublier nos gènes et nos pages web. C'est de cette évidence statistique que se sont emparés depuis quelques années les acteurs du marketing et des ventes, de la communication ou des métiers de l'information. On aime ainsi à expliquer le succès d'Amazon par sa couverture en produits aussi bien pour un public large adepte des mêmes modes (les *hubs* commerciaux que tout le monde s'arrache) que pour l'acteur local individuel à la recherche d'une rareté (les objets originaux et parfois inattendus que chacun peut collectionner, la *long heavy tail* de la courbe). Rien n'y échappe, depuis les degrés de contribution à Wikipedia (quelques contributeurs massivement impliqués et une foule de contributeurs occasionnels) jusqu'aux comptes Facebook. Mais, si l'on admet qu'il existe une couche intermédiaire, celle qui abrite sur le Web les agrégats de pages regroupées thématiquement, il faut alors couper la courbe non pas en deux, mais trois parties distinctes !

Image 5-4

Retrouver la zone d'agrégation dans la distribution des sites en loi de puissance.

Méthodologiquement, le principe est d'une grande conséquence. Pour un système d'information comme le Web, disons, un système documentaire *distribué* et *ouvert*, cette zone intermédiaire – où la courbe change de direction – est celle où s'exerce le principe de corrélation contenu-structure, là où se situent les *topical localities* (localités thématiques²). Si l'on examine les pages web situées plus haut dans la courbe, on trouvera des sites généralistes ou des plateformes type moteurs de recherche ou réseaux sociaux. Générant beaucoup de liens, ils n'ont pas de coloration sémantique particulière – hormis le fait qu'ils peuvent (presque) accueillir tous les types de contenus. Jusqu'à Twitter qui fonctionne comme une pure couche dynamique de *signalement* de contenus. Si l'on se saisit de la *long heavy tail* – la queue statistique du phénomène –, on trouvera des myriades de pages ou de sites web à peine reliés au reste du réseau (quelques liens seulement) connectées le plus souvent à une ou quelques *authorities* locales d'un seul agrégat.

2. F. Pfaender, M. Jacomy et G. Fouetillou, "Two Visions of the Web: From Globality to Localities", dans *Proceedings of IEEE ICTTA06*, 2006.

Je me demande parfois si l'attention actuelle portée autour du concept du *6 degrés* (enfin, des diamètres réduits notamment dans les réseaux sociaux) ou celle portée à la courbe en *loi de puissance* (aussi bien dans la science que dans le marketing) ne masque pas, en réalité, ce principe des couches. Si l'on s'extrait quelques instants du monde de la science des données et du Web, on s'aperçoit rapidement que d'autres domaines opèrent cette sorte de tripartition statistique comme une loi gouvernant bien des phénomènes dans le transport, l'industrie, le commerce et la logistique. En logistique, on trouvera par exemple rangé sous la rubrique « méthode ABC » un principe d'optimisation des coûts liés soit à la production industrielle, soit à la gestion des stocks et des quantités. On y met en classe A les 10 % des articles qui représentent 60 % de la valeur totale du stock ou du nombre de clients, en classe B les 40 % des articles qui représentent 30 % de la valeur totale du stock ou du nombre de clients et, enfin, en classe C les 50 % des articles qui représentent 10 % de la valeur totale du stock ou du nombre de clients. Derrière cette distribution en 3 classes, il y a aussi un modèle d'approvisionnement et des stratégies de gestion associée à chaque classe de produit ou de clients. En classe A les quelques produits phare en flux tendu que s'arrache un grand nombre de clients, en classe C les milliers de produits qu'il faut garder en stock pour répondre aux demandes particulières et, entre les deux en classe B, les produits vendus *par lots* à des *communautés particulières* de clients. L'équivalent de nos agrégats de documents web.

Le *modèle topologique du Web en couches* représente une avancée importante au plan théorique dont on peut tirer quelques leçons pour le *design* des technologies web et pour les méthodes d'observation (voire de veille). La publication de l'article de Jon Kleinberg où il expose les principes de HITS correspond à une époque qui voit se développer la technologie des *focus crawlers*, une technologie fascinante conçue pour isoler des agrégats de documents web et les parcourir de façon systématique. La confrontation de notre modèle topologique du Web avec les technologies de l'information a constitué le moment central de nos aventures, un moment privilégié de la recherche technologique où se mêlent hypothèses et statistiques, concepts et *data*.

Les *web crawlers*, au sens général, sont un élément central des technologies type moteurs de recherche et sont indissociables de la structure du Web lui-même : en passant d'URL à URL, ou en suivant chacun des fils de la toile, ils permettent de reconstruire un plan local, voire général, de l'ensemble de la structure documentaire hyperliée. Si vous décidez d'explorer le Web sans navigateur ni moteur de recherche,

ils sont indispensables pour mener des explorations de façon autonome. Si l'on met de côté les aspects d'optimisation d'un moteur de recherche complet (comme les temps de réponse, les calculs de pertinence par rapport à une requête, etc.), on peut les mobiliser pour leur valeur expérimentale et exploratoire. Le *web crawling* est un processus qui suppose une infrastructure de stockage des données (au plus simple, ce sont des listes d'URL et de connexions hypertextes que l'on va convertir ensuite en structure de graphe avec ses nœuds et ses liens), mais aussi deux réglages nécessaires. Le premier paraît naturel : étant donné les capacités limitées d'un système de stockage des données (sans oublier l'espace de calcul nécessaire associé à leur traitement sous forme de graphe), il faut spécifier la *profondeur* du processus, autrement dit la distance maximale qu'il doit parcourir depuis son point de départ. Si l'on n'a pas oublié que le Web, malgré le nombre énorme de documents, est un réseau dont le diamètre est relativement court, ce réglage de la profondeur est en général très faible (2, voire 3). Ensuite, il faut assigner au *web crawler* un point de départ, voire plusieurs. Cela n'a rien d'un détail : si chacun se tourne spontanément vers Google ou son encyclopédie préférée pour y chercher des informations fiables et essentielles, par où commencer l'arpentage de la structure du Web comme réseau documentaire ? Personnellement, je pense que toutes les méthodes sont bonnes du moment qu'elles sont explicitées : piochons dans nos *bookmarks*, dans ceux des autres, et même dans les moteurs de recherche ou les réseaux sociaux, pour lancer le processus. Dans le premier modèle du Web en nœud papillon, le nombre de points d'entrée sur le Web était de 570, mais on n'a guère de précision sur la façon dont ils ont été sélectionnés. Ces points d'entrée dans la structure constituent une *seed* à partir de laquelle vont se développer différentes branches qui finiront, peut-être, par former un vaste réseau connexe où tous les éléments identifiés se trouvent reliés entre eux.

Parmi toutes les techniques de *web crawling*, la technologie et les méthodes de *focus crawling* sont particulièrement profilées pour explorer et indexer la couche médiane du Web, celle des agrégats de documents. Si l'on cherche une sorte de microscope dédié à leur observation, les méthodes de *web crawling* contraintes par un focus conviennent parfaitement, à condition de pouvoir maîtriser les difficultés techniques et méthodologiques qui leur sont associées. La principale difficulté consiste à *contraindre* le *crawler* à rester focalisé sur un ou plusieurs agrégats, tout comme on place précisément un satellite en orbite autour d'un astre faute de quoi il finit aspiré par d'autres trajectoires. Les techniques sont nombreuses et, pour chaque étude d'agrégat, il faut en associer plusieurs : interdire au *crawler* d'aller sur des URL définies (par exemple « blacklist » les URL du type *google.fr*, *wordpress.com*, *facebook.fr*... issues de la couche haute du Web, de la même façon que l'on lemmatise un texte en éliminant les « mots outils »), on peut limiter le processus en termes de navigation hypertexte (par exemple à une profondeur de 2 maximum) ou encore identifier

des points de départ déjà focalisés sur le thème poursuivi. À vrai dire, toutes les recettes pour empêcher le *crawler* de « remonter » vers la couche haute du Web sont pertinentes, même si le temps de l'expérimentation et des différents réglages de la technologie de *crawl* peuvent être longs (et même variables suivant les thématiques web observées). Une stratégie, clairement évoquée par Jon Kleinberg, consisterait aussi à « découper » dans un grand graphe web un *cluster* identifiable pour sa densité de liens internes et dont les URL seront les seules à être crawlées. Dans ce cas, l'idéal, évidemment, serait de posséder *par avance* une liste des meilleurs *hubs* du domaine et de régler la profondeur à 1 seulement, ce qui permettrait rapidement d'identifier toutes les meilleures *authorities*. Sans disposer de l'infrastructure d'un moteur de recherche généraliste et de ses grands graphes, chacun est condamné à bricoler localement avec la méthodologie... d'où on peut tirer quelques innovations intéressantes.

Car ces techniques de contrôle du déploiement des *crawlers* web ne suffisent pas toujours : le principal danger est de voir le processus déborder, voire se perdre, dans un ou plusieurs agrégats voisins (ce que les chercheurs d'IBM du projet CLEVER³ ont appelé *topical drifting*). C'est là qu'intervient le principe du *focus crawling*. En son principe, il repose sur l'alternance de deux processus : la navigation par liens et l'analyse du contenu des pages web rencontrées. En général, il faut entendre par « analyse du contenu » la production d'une liste (voire d'une table ou matrice de co-occurrences) de mots-clés, classés par exemple par degrés d'occurrence sur une page. Chaque *focus crawler* est fondé sur une formule d'hybridation des deux dimensions – *recherche des liens* et *identification des contenus*. Par exemple, une formule peu contraignante (large) permet de laisser le *crawler* naviguer et de faire « remonter » pour chaque agrégat rencontré des sortes d'ontologies (qui peuvent préfigurer une classification systématique des agrégats et de leurs thèmes). Une formule particulière consiste à intégrer une liste de mots-clés (pondérés si possible) et à intégrer au graphe seulement les pages qui possèdent tel ou tel mot-clé, ou tel score de mots-clés (en additionnant par exemple les scores et le degré d'occurrence). Cette dernière méthode est particulièrement efficace pour isoler un agrégat, et le revisiter régulièrement dans un processus de veille. Elle a aussi l'avantage de faire apparaître les frontières d'agrégat (de corpus) si l'on conserve en archive les URL exclues du processus de classification automatique des contenus (pages intégrées au corpus / pages exclues du corpus).

3. <http://www.research.ibm.com/labs/almaden/index.shtml>.

Image 5-5

Les frontières d'un agrégat peuvent apparaître avec un *focus crawler*.

En lançant nos premières expérimentations dès les années 2002-2003, nous empruntons là aussi des chemins déjà largement balisés par nos pairs américains... et parfois indiens. Venu de l'IITB (l'Indian Institut of Technology of Bombay, l'équivalent indien du MIT), Soumen Chakrabarti synthétise en 2002 l'ensemble des enjeux et des techniques de *focus crawling* dans son ouvrage central *Mining the Web: Discovering Knowledge from Hypertext Data*. Très pointu techniquement, l'ouvrage (comme les publications scientifiques de l'auteur) s'adresse à des experts ou des ingénieurs compétents en technologies web d'indexation. Si, de surcroît, vous maîtrisez les mathématiques statistiques, le livre fera votre bonheur. Étant notablement incompetent en matière de mathématiques, les méandres du codage des *focus crawler* m'échappaient largement. Je compris cependant grâce aux premiers chapitres de l'ouvrage les différentes façons de décliner le principe du focus, a *small-* ou *medium-scale* sur le Web, éventuellement de façon fédérative (plusieurs *crawlers* fonctionnant en communautés d'agents intelligents) et quelques principes de classification automatique par apprentissage qui permettaient d'affiner le focus en cours de processus (processus largement contrôlé ou supervisé par un expert). Mais l'importance de l'ouvrage de Soumen Chakrabarti apparut clairement pour les conclusions théoriques et méthodologiques qu'il tirait de ses expérimentations avec des *focus crawlers*. L'auteur insistait en particulier sur la valeur scientifique de l'observation du comportement des *focus crawlers* dans leur course sur le Web, traversant ici des zones de haute densité de liens hypertextes, et là des zones presque vides où seulement quelques liens hypertextes reliaient des documents épars. Dans une première étape, chaque zone de densité est pour lui l'occasion de s'interroger sur les contenus qu'elle agrège et, au-delà, sur la *web community* qui l'anime. Exactement de la même façon que

nous explorions certaines « thématiques » du Web, comme la controverse autour de l'Église de scientologie, selon le principe de *knowledge discovery* appliqué au Web tel que Jon Kleinberg l'avait décrit avec HITS. Soumen Chakrabarti ne parle pas d'ailleurs de « couche médiane » du Web, mais il assigne clairement sa dynamique d'agrégation aux activités multiples des communautés sociales qui les produisent (*discovering ou identifying web communities through linkage topology*). Tous les efforts techniques et méthodologiques qu'il développe se concentrent sur l'objectif de repérer, avec des ressources minimales, les meilleurs *hubs* et les meilleures *authorities* d'un domaine informationnel nourri par cette dynamique sociale. Il est donc conduit, comme l'avait fait Jon Kleinberg avec HITS, à inventer une sorte de métrologie documentaire en *web mining* où sont hybridées des mesures de distance hypertexte et de similarité sémantique. Le type de technologie dédiée par excellence à l'exploration de la couche médiane du Web, à la granularité transversale du système où *proximité* et *similarité* semblent résonner et se renforcer en un signal clairement identifiable.

Soumen Chakrabarti teste donc le comportement de ces différents types de *crawlers* en les déployant à différents « endroits » du Web, autrement dit en faisant varier ses points d'entrée dans le système (la *seed* des URL à partir desquelles le robot entame son travail d'exploration de lien en lien). Ses expérimentations ont pour premier objectif de démontrer toute l'efficacité d'un *crawler* de type « focus » comparé à un *crawler* « classique » qui suit les liens hypertextes sans produire d'indices sur le contenu des pages visitées. À l'inverse, un *focus crawler* intègre des règles de visitation ou des contraintes de parcours sur le Web. Par exemple, il ne se déplace pas à partir de pages peu pertinentes ou « hors sujet » ; en revanche, si deux pages pertinentes (A et B) en pointent une troisième (C), il va privilégier cette direction dans le graphe hypertexte du Web. Évidemment, l'hybridation méthodologique entre analyse du contenu et repérage des liens hypertextes peut être d'une grande complexité, surtout dans une perspective d'optimisation de recherche rapide et exhaustive de documents pertinents sur un sujet.

Il teste ainsi les deux types de technologies (*sans* et *avec* focus) en prenant les mêmes URL de départ pour « lâcher » les robots sur le Web, ces URL traitant pour la plupart du même thème (*cycling*). Comme le montrent les tableaux statistiques ci-dessous (a), le *crawler* sans focus se perd rapidement sur le Web, ne ramenant (*harvest rate*) que quelques documents pertinents, identifiés sur de longues distances où un grand nombre de documents sont crawlés. À l'inverse, le *focus crawler* reste concentré autour des URL de départ et ne se déplace que dans le périmètre restreint de l'agrégat, concentré sur le thème choisi et les pages pertinentes, identifiant rapidement les « meilleurs » documents (*authorities*) à partir des meilleurs *hubs*. Les indices statistiques sont sans appel (a).

Image 5.6

(a) Taux comparé d'extraction de pages pertinentes durant un *crawl* (en ordonnée le degré de pertinence estimée des pages extraites durant le *crawl*, en abscisse le nombre de pages extraites). À gauche, les résultats statistiques du *crawler* sans principe de focus, à droite, les résultats d'un *focus crawler* sur le même thème (*cycling*). (b) Répartition de 100 pages web analysées comme pertinentes en fonction de la distance en nombre de liens hypertextes établis à partir de plusieurs URL de départ. Dans cet exemple, la concentration des documents pertinents se situe, pour l'exemple de gauche, entre 2 et 10 « clics » d'un ensemble de points de départ (*seed*) et, pour celles de droite, de 9 à 15 « clics » : le taux moyen de détection de pages web pertinentes (*harvest rate*) en fonction « d'où commence » le *crawl* dans l'architecture granulaire du Web en agrégats.

Crédit : S. Chakrabarti, Mining the Web: Discovering knowledge from hypertext data, 2002. Morgan Kaufmann.

D'une certaine façon, le *focus crawler* reste dans le périmètre des points de départ, des URL pertinentes en termes de contenu (*focused pages* sur le thème *cycling*). Mais que se passe-t-il si les points de départ du *focus crawler* sont éloignés, dans l'univers hypertexte du Web, de l'agrégat recherché avec ses *hubs* et ses *authorities* ? Ou encore, comment se comporte le *crawler* si on modifie son focus (en lui demandant par exemple d'aller chercher des documents pertinents sur « *mutual funds* ») alors qu'il est dans cette zone du Web où l'on parle de « *cycling* » ? En somme, peut-il *se déplacer d'agrégat en agrégat* dans cette couche médiane, comme une sorte de navigation latérale ? Bien évidemment, un *crawler* classique qui fonctionne sans inférence particulière sur sa façon de naviguer devra peut-être

parcourir l'ensemble du Web, y compris sa couche haute pour trouver finalement le bon agrégat, s'il ne s'est pas perdu en cours de route dans une structure de centaines de milliards de liens ! Le *focus crawler* intègre quant à lui des règles de navigation qui le rendent plus performant, dans un parcours latéral qui peut être long, mais couronné de succès comme le montre le second graphique (b). Durant ce parcours, les chemins possibles sont nombreux, mais, depuis son point de départ, dès qu'il rencontre des pages inutiles (et, mieux encore, des *hubs* qui ouvrent sur des agrégats non pertinents), il peut abandonner la piste par élagage du graphe (*pruning*), faisant émerger des chemins spécifiques à emprunter et des tâches prioritaires à exécuter (*priorization*). Dans le cas de l'expérimentation (b), le *crawler* atteint l'agrégat sur lequel il se concentre alors à partir de 9 liens ou « clics » de distance, ce qui en théorie est un chiffre très important quand on a en tête une estimation générale du diamètre du Web autour de 12 ou 13. Mais, si l'on se rappelle que le *focus crawler* établit à chaque déplacement des inférences sur la nature des agrégats qui l'entourent, son cheminement peut alors ressembler à une sorte de tunnel (*tunneling*) qui le conduit à travers la couche médiane du Web.

Chakrabarti, Kleinberg et les chercheurs d'IBM du projet CLEVER avaient ainsi imaginé les propriétés d'une forme de technologie dédiée au Web et à sa structuration en agrégat. La notion de *territoire numérique* prenait ainsi corps et les processus innovants de recherche d'information sur le Web développés à l'époque en validaient le principe. La méthode du *focus crawling* permettait d'envisager des technologies et des méthodes dédiées à l'exploration de la couche médiane du Web, là où se croisent distribution des liens et des contenus en agrégats typiques. Quand on découvre ces contributions scientifiques (parfois très techniques pour qui n'est pas ingénieur), on devine rapidement l'intérêt, et peut-être l'émerveillement, des chercheurs devant le dévoilement d'une organisation de l'information qui n'était pas héritée des bibliothèques ou des systèmes d'information avec des catégories préconçues et logiquement distribuées en hiérarchies. On aurait dit qu'ils découvraient pour la première fois un territoire contributif dont l'histoire commençait en 1945 avec le projet Memex imaginé par Vannevar Bush.

Évidemment, la notion de *territoire numérique* fait encore débat aujourd'hui, ne serait-ce que pour les rapports ambigus entretenus avec la géographie et les territoires physiques. Et les débats peuvent se poursuivre autour de la cartographie d'information qui reste voisine, malgré tout, de sa grande sœur la cartographie physique avec sa longue tradition. Cependant, il m'apparaissait clairement que la voie ouverte par Jon Kleinberg et tant d'autres devait être considérée comme un appel au développement de nouvelles formes de technologies de l'information dédiées aux particularités de ce réseau des réseaux qu'est devenu le Web. C'était une façon efficace de réenchanter

la technique et les instruments d'exploration, une incitation à imaginer de nouveaux produits innovants braqués sur un processus de découverte des connaissances. Que des millions d'internautes contribuent massivement à travers la diffusion de milliards de pages web à faire émerger de telles formes d'agrégats avait quelque chose de fascinant, comme une sorte de *savoir non su* ou de connaissances partagées et implicites que la technologie des *focus crawlers* propulsait en plein jour.

* * * * *

La lecture de ces contributions scientifiques autour du principe d'agrégation contribua directement à cette sorte d'effervescence collective qui vous fait basculer, sans le savoir, dans la recherche technologique et l'innovation. Pour les jeunes ingénieurs qui m'accompagnaient, il devenait urgent de nous doter de nos propres instruments et d'asseoir encore un peu plus notre autonomie. La théorie des agrégats de documents web annoncée par HITS en 1997 et validée par les contributions de Soumen Chakrabarti (et d'autres) a constitué la clef de voûte intellectuelle d'une série de développements qui alimentèrent notre boîte à outils, et le point de départ de bien des aventures en termes de projets professionnels. Tout a commencé pour nous avec la conception et la mise en œuvre de notre outil phare, TARENTE, un instrument d'exploration et d'analyse des agrégats de documents web⁴. TARENTE était bâti autour d'un *crawler* que l'on pouvait équiper d'un focus sémantique et qui permettait d'observer en temps réel l'apparition à l'écran d'un agrégat via une solution de spatialisation type *force directed*. L'instrument était doté d'outils statistiques qui permettaient pour chaque page récupérée sur le Web de calculer une table de mots-clés co-occurents et classés par ordre statistique. Il permettait aussi d'activer la visualisation des pages web via un navigateur interne.

4. F. Ghitalla, E. Diemert, C. Maussang et F. Pfaender, "TARENTE: An experimental tool for extracting and exploring web aggregates", dans *Proceedings. International Conference on Information and Communication Technologies: From Theory to Applications*, IEEE, 2004.

Image 5-7

Copie d'écran de TARENTE.

Peu de temps après, d'autres ingénieurs se mirent à développer cette curieuse version de *Firefox* que fut le *Navicrawler*⁵. L'idée en était simple : on pouvait adjoindre à *Firefox* une série de *plugins* qui permettaient de suivre, d'archiver et d'analyser des parcours thématiques sur le Web sous forme de graphes enrichis. Le *Navicrawler* se présentait comme la *Rolls Royce* du contrôle qualitatif des *web data* pour la construction de corpus. Il intégrait un *crawler* hypertexte à partir des URL rencontrées (elles-mêmes intégrables ou non au corpus), la possibilité de les taguer à la volée à partir de labels choisis à l'avance, des calculs statistiques sur le graphe produit par la navigation, des fonctions d'import/export des données générées et une connexion en temps réel avec *Gephi*... Il ne m'en fallut pas plus pour imaginer l'infaisable quelques mois auparavant : partir à plusieurs à la recherche d'un agrégat de documents web et visualiser en temps réel les multiples chemins se suivre ou se croiser ! Une fois l'idée lancée et un court travail de préparation, le dispositif que nous avons imaginé a été présenté pour la première fois au printemps 2009 au médialab de Sciences Po : pendant que je parle à la tribune (accompagné de Mathieu Bastian), s'affiche derrière moi un graphe en expansion continue où figurent des URL ayant pour thème générique « la loi Hadopi »⁶.

5. <http://webatlas.fr/wp/share/navicrawler/Guide%20m%E9thodo%20NC%202007.pdf>.
6. https://www.youtube.com/watch?v=BEk03i8rj_w&list=PL6CAD477CD8A6CCA5.

Image 5.8

Inauguration du médialab de Sciences Po.

Les URL sont représentées sous la forme de nœuds de couleur correspondant à trois recherches réalisées simultanément par trois ingénieurs situés en marge de la salle de conférences (sur trois systèmes d'exploitation : Linux, Windows, Apple et en utilisant trois moteurs de recherche différents). Chaque nouvelle URL validée par l'un des trois ingénieurs est automatiquement intégrée au graphe général via l'identification de tous les liens sortants qui la relie au reste de la toile. Cette *performance technologique*, comme on pourrait le faire dans le monde de l'art contemporain, démontrait qu'il était possible de déployer sur le Web une veille collective, avec un procédé centralisé de visualisation, en temps réel, dédié à l'exploration d'un agrégat dont les URL étaient liées et très calibrées thématiquement autour des débats sur la loi Hadopi. Comme avec un filet de pêche, nous venions de ramener en quelques dizaines de minutes l'un des agrégats les plus dynamiques à l'époque sur le Web francophone. Le principe peut être poursuivi, par exemple en déployant sur l'agrégat isolé (donc un corpus de sources pertinentes sur un sujet et reliées entre elles par es liens hyper-textes) un *tracker*, autrement dit un *crawler* à qui l'on demanderait de travailler en continu sur le corpus pour évaluer tout changement dans le temps (par exemple inspecter de nouvelles ressources).

Une fois admis le principe d'une agrégation continue des ressources web dans une couche médiane, on peut ainsi commencer à forger des instruments qui seront dédiés à son exploitation. À l'époque, ce que je n'avais pas prévu s'imposa à la plupart de mes jeunes associés : pourquoi ne pas inventer des services et des entreprises qui exploiteraient toutes ces nouvelles propriétés du Web, donc des informations en réseau ? C'est ainsi qu'après leur participation aux premières expéditions du Web, Alain Le Berre et Guilhem Fouetillou fondent en 2004 la *start-up* RTGI à laquelle nous sommes nombreux à avoir été associés comme cofondateurs. L'une des colonnes vertébrales de l'entreprise (devenue aujourd'hui Linkfluence) est constituée par la découverte, l'exploitation et l'entretien de tous ces agrégats autour desquels se

regroupent les internautes dans de nombreux domaines (loisirs, politique, sports, économie). Avec les instruments adéquats et des méthodes à la fois quantitatives et qualitatives (avec les chargés d'étude qui sélectionnent et analysent les ressources), la société peut proposer des services inédits de veille sur le Web, en particulier sous l'angle de la *e-réputation* et du *tracking* des phénomènes émergents ou des tendances. En deçà de l'activité commerciale, la société a tenu à maintenir depuis sa fondation son dispositif de veille sur l'ensemble des agrégats de documents web (utiles à ses services) dans un système d'*écosphères*, comme le ferait un cultivateur soucieux de ses champs. C'est là, dans cette couche médiane du Web où liens hypertextes et contenus se concentrent et se répondent, que Linkfluence surveille ses territoires de prédilection. Pour être franc, il ne m'est jamais venu à l'esprit, au départ de nos premières expéditions sur le Web et les réseaux d'informations, que l'on puisse en tirer un projet de création d'entreprise de cette ampleur. D'autres projets ont éclos par la suite issus d'un même esprit : Linkurious (Sébastien Heymann) et l'Atelier Iceberg (Thomas Busson, Thomas Dupeyrat et Alexis Jacomy), ou Gephi (notre application favorite avec Mathieu Bastian en *lead developer*). Rétrospectivement, je prends conscience en écrivant ces lignes combien, à côté de la démarche scientifique, l'entreprise doit être considérée comme une forme authentique de validation d'un modèle, notamment par la mise en place de services nés directement de l'expérimentation technologique. Mais je me souviens aussi qu'à l'époque, hormis quelques-uns de mes collègues de l'UTC, personne n'a cru à ce type de démarche émergente, nourrie de surprises et d'imprévus, en pariant sur la production rapide de nouvelles idées imprévisibles trois mois auparavant. En serait-il autrement aujourd'hui dans l'univers de l'innovation administrée, des programmes prévisionnels ou des plans quadriennaux ?

Chapitre 6. Des *data* ciselées

« J'ai vu un ange dans le marbre et j'ai seulement ciselé jusqu'à l'en libérer », Michel-Ange.

L'identification d'une couche médiane sur le Web – ce qui allait devenir *la théorie des agrégats* – et nos premières expérimentations sur les *web data* ont profondément renouvelé les façons de nous représenter le réseau. Le principe des couches venait éclairer d'une façon inédite tout aussi bien le *6 degrees* (ou les diamètres moyens relativement courts des réseaux distribués) assignable à l'existence d'une couche haute hyperconnectée que la propension évidente des pages web à s'agréger en *clusters* remarquables, les agrégats dans une couche médiane où entraient en résonance liens et contenus. Cela nous rassurait : les *masses*, la *diversité* et la *dynamique* des *web data* étaient des contraintes que l'on pouvait désormais envisager de dépasser. Le principe d'une corrélation forte dans les corpus de pages web entre *proximité* hypertexte et *similarité* ou *complémentarité* sémantique permettait d'entrevoir de nouveau, comme dans une bibliothèque, des capacités d'exhaustivité et de recensement systématique de ressources pertinentes. Pour peu que l'on s'y penche, l'approche avait tout d'une sorte de méthode dédiée à des opérations de suivi ou de veille sur le Web. En revanche, nous n'avions pas la logistique suffisante pour explorer à grande échelle cette couche médiane et dresser un premier catalogue d'agrégats que l'on aurait pu traiter eux-mêmes sous forme d'un « métagraphe ». Après un échec auprès de l'ANR pour financer ce type de besoin et auprès de structures diverses (dont la région Picardie et mon université), il apparut évident à tous les membres de la jeune équipe qu'il fallait nous doter nous-mêmes d'une infrastructure d'exploration, et quelques serveurs OVH firent l'affaire dans un premier temps ainsi que quelques ordinateurs qui engloutirent nos économies personnelles.

Ce fut une époque formidable quand j'y repense : en marge des cadres officiels de la science, il a fallu imaginer différents scénarios afin de poursuivre notre démarche exploratoire. Les contraintes de la situation et le dynamisme de jeunes ingénieurs inventifs nous conduisirent à oser parier sur des voies encore inédites pour certaines. Quelques-uns parmi eux fondèrent la *start-up* RTGI (d'après un texte prospectif que j'avais écrit aux premiers jours de l'année 2000), devenue Linkfluence

depuis. D'autres préparaient les différentes premières versions de Gephi dans un cadre *open source*¹. Cette période mouvementée nous permit aussi de tester quelques nouvelles idées, surtout en termes de méthodologie et d'instrumentation. Nous étions imprégnés de l'idée que le succès de la démarche de Kleinberg et de ses collègues de Palo Alto ne tenait pas aux puissances de calcul mises en œuvre (ou de la taille des graphes du Web), mais à l'intelligence intégrée aux algorithmes comme HITS, intelligence que l'on pouvait désormais embarquer dans une flotte de *crawlers* sur le réseau. En attendant de bénéficier d'une infrastructure technique dédiée à l'exploration à grande échelle des agrégats, nous nous sommes focalisés sur les aspects qualitatifs de nos méthodes et de nos instruments. Il nous fallait donc trouver un terrain d'expérimentation assez ouvert pour intégrer différents agrégats et où nous aurions pu tester de nouvelles pistes pour traiter les données. Au-delà d'une structure d'agrégat finie et localisée, il nous fallait comprendre comment s'organisaient les agrégats *entre eux*, de voisins en voisins, de façon à entamer une description minutieuse de la couche médiane du Web.

Mathieu Jacomy m'accompagnait, alors qu'il développait par ailleurs une sorte de navigateur enrichi, le Navicrawler, qui permettait d'intégrer à Firefox des fonctionnalités de cartographie des parcours de navigation à base de graphes². Mathieu s'intéressait au principe de la *navigation enrichie* ; il s'interrogeait par exemple sur l'utilité de taguer à la volée les sites web, de construire des corpus d'URL avec des batteries de descripteurs ou encore de traduire sous forme de graphe notre présence sur le Web en analysant les *logs* des utilisateurs. Il s'était ainsi tourné vers l'étude de ces allers-retours contenu-structure (des mots aux liens) et structure-contenu (des liens aux mots) que permettait le Navicrawler, mais de façon contrôlée et manuelle, presque artisanale. C'était une autre façon de visiter les agrégats de documents web. Loin des grands corpus de *web data* et des technologies automatiques d'extraction comme les *focus crawlers*, Mathieu aimait à produire ce genre de documents qui accompagne une méthode de construction d'un corps d'URL via une sorte d'arpentage systématique :

1. M. Bastian, S. Heymann et M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks", dans *Proceedings of the Third International ICWSM Conference*, 2010.

2. <http://webatlas.fr/wp/navicrawler/>.

Image 6.2

Méthode de sourçage d'un corpus web avec le Navicrawler.

À force de multiplier les expériences de construction contrôlée, quasi minutieuse, de corpus d'URL sur une foule de sujets sociétaux (les communautés *pro-life* ou *pro-choice* dans la polémique en ligne sur la contraception, les organisations de la solidarité internationale, la culture scientifique et technique sur le Web³...), il en était arrivé au point de synthétiser les différents aspects d'une exploration systématisée d'un agrégat :

3. F. Pfaender, M. Jacomy et G. Fouetillou, "Two Visions of the Web: From Globality to Localities", dans *Proceedings of IEEE ICTTA06*, 2006.

Image 6.3

Images de la méthodologie de *crawl* avec le Navicrawler.

Ce n'est que plus tard que nous avons réfléchi ensemble aux implications théoriques de notre démarche et aux enjeux qui s'y trouvaient liés pour les sciences humaines et sociales. Pour le moment, il fallait trouver l'occasion propice pour nous lancer dans ce genre d'expédition cartographique dans les *web data* qui nous amusait tant.

Nous n'avons pas attendu longtemps : le ministère de la Recherche (la cellule Point de contact national pour les questions européennes de la recherche – PCN) nous contacta pour réaliser une étude exploratoire de vaste ampleur sur les rapports « science-société »⁴. L'objectif général du projet, SISmap pour Science and Society Mapping, consistait à cartographier « sur le Web francophone les zones d'interaction entre acteurs et institutions de la science et acteurs et organisations de la société civile ». Le « focus » thématique nous paraissait suffisamment large pour englober une foule de problématiques potentiellement traitables sous forme d'une série d'agréats de documents web : développement économique, développement durable, sciences citoyennes, énergies, transport, nature, nourriture, chimie, nanotechnologies, sciences du vivant, espace, biodiversité, nucléaire, cosmétiques...

4. Sophie Tocreau et Martine Roussel.

Nous savions déjà par expérience, pour avoir étudié de près les thèmes controversés de l'avortement, de l'Église de scientologie, du logiciel libre ou du « mouvement des sans-papiers », que l'agrégation de contributions en ligne était particulièrement remarquable chez des acteurs engagés politiquement ou philosophiquement. Le dispositif imaginé pour SISmap était à notre connaissance tout à fait inédit. Il reposait sur l'hybridation de deux éléments qui intervenaient à tour de rôle. Le premier était constitué de quatre personnes (deux expertes du ministère, spécialistes des relations science-société, assistées de Mathieu et moi-même). Le rôle de cette cellule (qui se réunissait régulièrement) était d'opérer sur les données des tâches de sélection (pertinent/hors sujet), autrement dit de déployer à chaque étape du processus des activités qualitatives d'interprétation et de sélection à l'image de ce que ferait un *focus*, mais ici manuel et collectif. À chaque étape, la cellule permettait donc de valider et de contrôler l'expansion du corpus, en particulier en s'appuyant sur des épreuves imprimées de nos graphes autour desquels s'effectuait le travail d'analyse collective.

L'autre élément, plutôt quantitatif celui-là, était constitué par Examiner. Examiner était un prototype expérimental dont l'ossature reposait sur des scripts exploitant l'API Exalead. Exalead (exalead.fr) est un moteur de recherche français qui contient 16 milliards de pages en base, issu d'une équipe d'experts de haut vol en matière de *search* (fondé en 2000 par d'anciens développeurs d'Altavista, François Bourdoncle et Patrice Bertin). Toujours prête à se lancer dans de nouvelles aventures, l'équipe de l'époque (au sein de laquelle œuvrait Hugo Zanghi aujourd'hui à Linkfluence) accepta volontiers l'autorisation d'exploiter leur base d'indexation de pages web. Notre objectif résidait, en particulier, dans l'extraction des relations statistiques entre des termes ou des expressions (en langue française) et qui constituent souvent des indicateurs de contenus pertinents pour un travail à grande échelle. Nous avions besoin d'accéder à des calculs statistiques simples, mais massifs qui auraient permis de dégager des *patterns* de mots-clés pertinents pour notre étude. S'il revenait à la cellule d'experts de se prononcer sur l'inclusion ou non de certains termes ou expressions, les liens entre termes étaient extraits à partir de leur degré de co-occurrence sur le Web, en ce cas dans les bases d'Exalead. Un terme A (par exemple *pollution*) est lié à un terme B (*polymère*) parce qu'ils apparaissent ensemble dans la même URL ou la même page, et selon une certaine intensité (par exemple, *pollution* 8 occurrences et *polymère* 6 occurrences dans la même page). Les termes étant par ailleurs liés par le même phénomène à d'autres termes, il devenait possible d'étendre le graphe de façon presque illimitée, en espérant faire apparaître des zones de concentration typiques, des « thèmes » si l'on veut. Les grammaires possibles d'attribution de scores sont très nombreuses en *text mining* et l'on peut raffiner l'analyse avec des traitements plus avancés, par exemple avec des formules d'analyse sémantique. Dans le cadre de SISmap, nous n'avions besoin que de statistiques simples dans notre arpentage du

Web, en espérant voir apparaître des zones de concentration de vocabulaire typique d'une agrégation après différentes phases d'analyse et de validation qualitative.

Ce sont les travaux de Kleinberg sur les agrégats de documents web qui nous ont incités, comme tant de chercheurs et d'ingénieurs, à tester différentes formules d'association où alternent phases de traitement des contenus et phases de traitement de la structure hyperliée. Les techniques de *knowledge discovery* exposées par Kleinberg et ses collègues consistent avant tout à exploiter la structure hyperliée pour en explorer les contenus dans une seconde phase. *Detecting web communities through linkage* : c'est tout le sens de HITS où le *topic search* se trouve induit (*induced*) par la distribution des liens (*hypertext*). La voie que nous avons empruntée est inverse et complémentaire pour détecter des agrégats sur le Web : nous voulions commencer *d'abord* par les mots pour étudier leurs concentrations typiques autour de thèmes controversés. Seulement *ensuite* viendraient l'analyse des pages et la distribution des liens hypertextes. Cependant, quand nous avons mis en place notre dispositif, rien ne garantissait que nous allions pouvoir nous aussi approcher un processus d'agrégation de mots et de liens avec cette méthode et, qui plus est, à grande échelle. Autrement dit, un test grandeur nature pour ce qui était devenu pour nous la « théorie des agrégats ».

Image 6.4

Architecture d'un dispositif de test de la théorie des agrégats.

Pour entamer le processus d'extraction des données Exalead, l'équipe d'experts du ministère a d'abord identifié 90 termes ou expressions qui ont servi de points de départ (l'équivalent d'une *seed* d'URL pour un *crawl* hypertexte sur le Web). De là, les données Exalead ont montré sous forme de graphe de mots-clés les liens de co-occurrence de chacun de ces termes avec d'autres, en grand nombre. L'équipe a sélectionné les expressions les plus pertinentes pour notre travail (*i.e* les expressions redondantes comme « matière de développement durable » issue de « développement durable », soit hors sujet par homonymie le plus souvent, *i.e* « Ocean Eleven » ou « Dany Ocean » issues de « océan »). Le graphe initial se voyait donc amputé de certaines branches, tandis que celles retenues comme pertinentes commençaient petit à petit à se rejoindre et à se connecter. En deux phases d'un long travail d'analyse manuelle, 940 termes de départ ont été retenus pour l'extraction finale de 6 760 expressions pertinentes et validées manuellement, associées entre elles par 71 744 liens d'association pondérés en fonction du nombre de co-occurrences dans les pages web, et fortement clusterisées.

Indéniablement, même sans avoir visualisé en détail le graphe, tous les indicateurs statistiques montraient que l'on venait d'arpenter un territoire unique (connexe) avec une morphologie interprétable (des hiérarchies, des *clusters*). Afin de vérifier que nos termes étaient bien issus de pages web pertinentes ou représentatives (au moins globalement), 26 787 d'entre elles (parmi plus de 240 000 mobilisées pour nos calculs) ont été conservées en base et restaient accessibles à tout moment depuis l'interface d'Examiner. Examiner était un dispositif de gestion de l'API Exalead d'investigation pluggé sur les données archivées par la société. Examiner tenait du bricolage méthodologique et technique, mais permit de produire un graphe général de la structure au format Gephi.

Projetés avec l'algorithme Force Atlas depuis Gephi, les 6 760 termes et expressions co-occurents ont été regroupés en 113 groupes (*clusters*) déterminés par différents seuillages statistiques. Un nombre de termes sans attribution (entre deux *clusters*) furent enfin intégrés après examen du groupe d'experts. Les liens, de plus ou moins grande intensité, indiquent des degrés de co-occurrence de termes *entre les clusters* (plus le lien est épais entre deux *clusters*, plus ils partagent de vocabulaire commun). Ce travail statistique et la projection spatiale confirmaient les effets de regroupement thématique, mais aussi d'agrégation plus ou moins intense dont dépendent les différentes grosseurs des nœuds. Les 113 *clusters* ou groupements thématiques ont été labellisés manuellement et proposent une distribution des 6 760 termes retenus.

170 mots Génétique	62 mots Agriculture biologique et produits bio
163 mots Préservation de la nature et de la biodiversité	62 mots Champs et rayonnements électromagnétiques
161 mots Commerce équitable	62 mots Risques, prévention et sécurité sanitaires
146 mots Développement durable et environnement	61 mots Accès aux soins et système de santé
142 mots Gaz à effet de serre et pollution de l'air	60 mots Découverte de la nature
139 mots Maîtrise de l'énergie	58 mots Urbanisme et territoire
137 mots Connaissance et domaine scientifique	58 mots Secteur agricole
127 mots Domaine de l'aide humanitaire	57 mots Éthique
127 mots Culture scientifique et technique et pédagogie des sciences	57 mots Transport routier et ferroviaire
126 mots Production, collecte, tri, traitement des déchets, traitement des eaux	56 mots Les différents types d'acteurs et partenaires
118 mots Chimie, laboratoires et universités	56 mots Secteur pharmaceutique et domaines thérapeutiques
116 mots Température et pression, mesure et variations	55 mots Qualité et distribution de l'eau
115 mots Aliments et alimentation	53 mots Mouvement social et alternatives au libéralisme économique
115 mots Controverse sur les OGM	53 mots Secteur financier
114 mots Biotechnologies, innovation et valorisation de la recherche	50 mots Acteurs politiques et écologie
113 mots Démocratie directe : participation citoyenne, crise de confiance	50 mots Risques d'infection et services hospitaliers
113 mots Cosmétiques	50 mots Centrales et production d'électricité
105 mots Construction européenne, Europe politique	49 mots Biologie et laboratoires
45 mots Loisirs et sports de nature	48 mots Les différents types de transports
43 mots Secteur maritime	47 mots Engagement des jeunes et secteur associatif
43 mots Différents ministères et ministères	46 mots Histoire et histoire des sciences, sciences politiques et géographie
43 mots Exposition aux risques	37 mots La citoyenneté et l'égalité des chances
42 mots Jardins	37 mots Grippe aviaire
42 mots Pêche	37 mots Gestion locale de l'eau
41 mots Cycle de l'eau	37 mots Sciences de l'ingénieur, matériaux et mécanique
41 mots Nucléaire militaire	35 mots Lutte contre le SIDA
41 mots Espèce humaine	34 mots Pauvreté
40 mots Personne humaine, dignité humaine	33 mots La pollution
40 mots Espèces animales	32 mots Le système biologique
40 mots Politiques publiques et choix scientifiques	30 mots Marché du gaz
40 mots Négociations du commerce international	30 mots Fête de la science
40 mots Prévention santé et sécurité du travail	29 mots Secteur aérien
39 mots Dette du tiers-monde	28 mots Peinture au plomb et amiante
39 mots Crise alimentaire	28 mots Les problèmes de poids
39 mots Services des Eaux	28 mots Coopération scientifique
39 mots Physique	27 mots Substances dangereuses
38 mots Populations et amélioration des conditions de vie	26 mots Gouvernance
38 mots ONU et droits de l'homme	26 mots Maladie de la vache folle
100 mots Domaine spatial et planète Terre	25 mots Sang
95 mots Filière nucléaire et stockage des déchets radioactifs	25 mots Entreprises artisanales
90 mots Milieu marin	24 mots Risque et population
90 mots Problématique et domaines du patrimoine	24 mots Carburants
	23 mots Grandes conférences
	23 mots Surveillance

Image 6.5

Réseau des 6 760 expressions pertinentes associées entre elles par 71 744 relations de co-occurrence.

Réduite aux seuls 113 groupements thématiques – 113 agrégats –, la carte affichait une forme curieuse en étoile avec des « branches » (ou des « filaments » comme les appelle Mathieu Jacomy). En bas ont été affichés quatre petits *clusters* thématiques déconnectés de la vue principale, aucune co-occurrence de termes ne le permettant. Peut-être s'agit-il d'un effet de notre méthode de sélection des termes pertinents pour notre étude, ou bien d'un manque éventuel de données supplémentaires. Cependant, je notais à l'époque que ces *clusters* déconnectés concernaient surtout l'univers politique, l'Europe ou le rôle des ministères... La carte produite ne possède pas de cœur dense et hyperconnecté, comme dans de nombreuses cartographies que nous avons produites. Ici, une distribution circulaire centrale égrène différentes thématiques dont certaines servent de pivot aux filaments.

Image 6.6

113 groupements thématiques d'expressions calculés par seuillage statistique.

À droite, c'est par exemple le cas de la thématique *Aliments et alimentation* : c'est à la fois l'un des *clusters* majeurs (comme *Développement durable et environnement*, *Génétique* ou *Préservation de la nature et de la biodiversité*), mais aussi le point de départ d'un filament qui nous fait passer par *Cosmétiques* ou *Sol et végétaux* pour s'achever sur l'univers de *l'agriculture* et du *commerce équitable*.

Image 6.7

Zoom sur le filament de groupes attaché au reste du réseau par le thème de l'alimentation.

L'exploration de chacun des 113 *clusters* de mots-clés co-occurents était rendue possible via Examiner : pour chacun d'eux un nuage de mots était associé, et chacun d'eux cliquable pour accéder aux documents source où ils apparaissaient.

Image 6.8

Le groupe d'expressions « Pétrole et marée noire » tel que visualisé dans le poster final du projet SISmap comme un nuage de mots.

La version imprimable de la carte contenait au final les 113 thèmes, les 6 760 termes associés ainsi qu'une partie des 71 744 liens de co-occurrence (ceux qui permettaient d'apercevoir les connexions inter-clusters dans le réseau de co-occurrence). Nous étions au bout des capacités de nos machines et de nos applications ! Affichée sur un mur, elle permet d'apercevoir le *pattern* central circulaire ainsi que les filaments, mais aussi, en se rapprochant de chaque nuage de mots-clés, de l'impressionnante précision des termes pour chacun d'entre eux

Image 6.9

Le poster final du projet SISmap.

Le foisonnement apparent des termes disparaît rapidement dès l'examen attentif des premiers *clusters* thématiques, comme si la forme globale de la carte leur imprimait une forme raisonnée de classement avec ses hiérarchies générales et ses complémentarités locales. Cette forme générale en étoile et ses articulations remarquables de thématiques ne laissaient pas de nous interroger : à quoi pouvait-on en attribuer le principe ? Ces formes d'*ontologies ascendantes* (puisqu'elles venaient du Web lui-même, fruit du travail continu des millions d'internautes) avaient-elles donc une forme jusque-là inaperçue ? D'où venait une telle organisation ?

Nous étions encore une fois confrontés à la délicate question de savoir à quoi assigner ce *pattern* fascinant : aux effets de notre méthode d'extraction des données, de nos instruments de spatialisation ou bien à un principe statistique robuste et pérenne sur le Web ? Étant donné les milliers de sites web d'où provenaient nos mots-clés, et le fait qu'ils étaient reliés entre eux pour la plupart (nous pouvions les consulter à tout instant via Examiner), il ne faisait aucun doute que nous avions cartographié une région du Web particulière où se concentraient par agrégation des contenus et des acteurs. Une fraction du Web, francophone et militant, où semblaient s'enchaîner les thématiques de controverses par complémentarité

(alimentation – cosmétiques – produits biologiques – agriculture biologique), mais aussi par la présence d'acteurs impliqués dans plusieurs controverses (France Nature Environnement, José Bové, Greenpeace France...). Puisque le *pattern* révélé par ForceAtlas 2 semblait si fort, alors pourquoi ne pas l'exploiter pleinement à travers un dispositif original ? À l'évidence, cette structure en étoile avec ses filaments pouvait être exploitée en la réduisant sous la forme d'une boussole pour naviguer dans les masses de pages web qu'Examiner avait identifiées et archivées. On aurait pu aussi associer à cet instrument cartographique d'aide à la navigation des listes de mots-clés importants : en cliquant sur un terme dans une liste adjacente, sa position aurait pu « s'allumer » sur la carte ou bien ouvrir une liste de sites associés.

Les idées s'enchaînaient et c'est ainsi qu'est né l'un de nos projets phare : un moteur de recherche dédié aux controverses scientifiques et techniques sur le Web. Nous lui avons donné un nom de projet : Réseaux, Traces, Controverses. Puisqu'il était possible de dresser la cartographie à peu près complète des controverses sur le Web (disons, des *sujets* de controverses sociales) à un instant t , qui plus est adossé à un système pérenne comme Exalead.fr, il devenait concevable de la reproduire à différents moments. Par exemple, chaque année, on aurait pu dessiner une morphologie socio-informationnelle des controverses (acteurs, organisations, arguments), les recenser sur le Web francophone (via une typologie) et en dresser la cartographie en fonction de leur « degré » d'éloignement ou de proximité, de différence ou de similitude. On aurait pu travailler sur les modalités d'interactions probables entre de nombreux *patterns* documentaires associés à chacune d'elles, en profondeur ou localement. Nous imaginions pouvoir nourrir le moteur d'une forme distribuée et collective d'expertise qualitative selon les mêmes procédures méthodologiques que nous avons suivies. L'une de nos idées favorites consistait à imaginer que, le temps passant, nous aurions pu produire des sortes d'indices ou d'*indicateurs de controverse* définis qualitativement (savoir d'expertise sur les controverses) et simultanément mobilisables pour des processus de traitement quantitatif sur de grandes masses de données documentaires (en particulier extraites du Web). Le *moteur de controverses* aurait ainsi pu servir de laboratoire, ou de base expérimentale pour étudier comment naissaient les controverses, à partir de quels arguments et de quels acteurs, comment elles évoluaient, disparaissaient ou fusionnaient pour générer de nouvelles formes. Pour couronner le tout, notre première carte semblait nous procurer un espace original d'exploration ou de navigation dans les quelque 27 000 pages web archivées et classées, aisément traduisible en forme interactive et que l'on aurait pu associer à des instruments plus classiques de recherche d'informations comme des listes générées à partir de requêtes ou une classification préétablie. Pour bien des raisons, indépendantes de notre volonté, ce moteur n'a jamais vu le jour. C'était pourtant l'un des premiers projets à prétendre hybrider les sciences de l'ingénieur

avec les sciences humaines et sociales dans l'étude et le suivi longitudinal des controverses scientifiques et techniques qui constituent l'une des formes les plus évidentes de la dynamique du social sur le Web.

Le ciselage des *data* auquel nous avait contraints notre recherche de précision et de pertinence fut un travail de longue haleine, mais ô combien instructif. L'étude nous avait poussés à imaginer plusieurs voies d'hybridation des procédés automatiques d'extraction de *web data* à grande échelle et travail qualitatif de sélection. Le prototype Examiner incarnait l'une de ces voies possibles. Mais, au-delà des instruments et de la technologie, au-delà même des *data* et des controverses, le plus instructif pour moi fut l'effort porté sur la construction d'un corpus cartographiable. Toutes les opérations que nous avons menées ont été des moments délicats et riches d'enseignement sur la construction de la carte comme *machine logique*. La visualisation de notre *pattern* « en étoile », après des mois d'effort, peut fasciner par son pouvoir visuel et révélateur. Mais l'émerveillement devant la carte fonctionne d'autant mieux qu'il s'agit d'une machinerie complexe que nous avons mis longtemps à maîtriser, comme un savoir-faire d'artisan cartographe. Tout est prêt pour que le regard du spectateur puisse appréhender des formes globales, mais aussi se perdre dans les détails en s'approchant du poster. De l'imprimé à la version interactive, il n'y a donc qu'un pas : pour produire un poster exhaustif, l'alignement des différents filtres nécessaires à la production de la vue est déjà en place et maintes fois testé.

En cartographie d'information, les pratiques d'atelier ouvrent donc directement sur des pistes d'innovations technologiques, mais à condition seulement que les procédures de traitement de l'information soient nourries d'une forme d'expertise. Notre cellule collective d'évaluation des versions successives de la carte mêlait des compétences en ingénierie et l'expertise des chargés de mission « culture et société » du ministère ce qui, en amont, a conditionné l'intérêt de la carte, aussi bien en précision qu'en étendue.

Image 6.10

Participants du projet SiSmap en cours
d'annotation des cartographies du projet au format papier.

On doit la vue en étoile à l'algorithme ForceAtlas 2 de Gephi, mais aussi, plus profondément, aux formes d'hybridation des compétences d'ingénierie et d'expertise. A priori, le premier est spécialiste des traitements quantitatifs au sens large, depuis les requêtes via l'API Exalead jusqu'aux bases de stockage des données et les calculs associés. Le second, l'expert, est plutôt spécialiste de la validation qualitative avec tout le travail d'inclusion/exclusion pour chacun des milliers de termes analysés. Mais, le plus souvent, le processus de coopération débouche sur des formes dynamiques d'hybridation. Ainsi, l'expert est appelé à fournir à l'ingénieur les critères selon lesquels extraire les termes ou en traiter les propriétés. Inversement, l'ingénieur peut relever l'apparition de nouveaux termes ou de nouveaux *clusters* de termes sur lesquels l'expert doit alors se prononcer. Pendant que l'un est garant des identités thématiques, l'autre est le garant des unités calculables. À regarder l'équipe se pencher sur les impressions papier successives de la carte, il apparaît clairement qu'il y avait là de quoi développer tout un travail précis de description des différentes modalités de coopération entre l'expert et l'ingénieur dans les formes de contrôle de production du corpus. C'est là l'un des secrets de la R&D à l'échelle industrielle des systèmes et des services web innovants. Mais c'est aussi là, dans le ciselage des *data*, que l'on cherche à embarquer de l'intelligence à travers les contributions d'experts. Je pourrais aussi le dire des chercheurs en sciences humaines et sociales : les modèles de temps, de sociétés, d'espaces, d'idéation ou d'interaction qu'ils produisent depuis longtemps constituent un formidable capital d'expertise potentiellement mobilisable. Mais cela supposerait une révolution des hommes et des services dans les organisations. Dans les universités, les centres de recherche ou les services transversaux

de grandes entreprises, les DSI (Directions des Systèmes d'Information, en général « quasi bunkerisés ») se trouvent souvent à l'opposé dans les bâtiments du service de veille, de la bibliothèque, du centre de documentation ou des services chargés des études. Le décroisement paraît pourtant nécessaire à l'évolution des technologies elles-mêmes et notre modeste atelier de *design* participatif *et* du corpus *et* de la technologie aura défriché une voie parmi d'autres du passage des masses de données à l'information pertinente, le passage du *big* au *smart data*.

Chapitre 7. Le Web Datarium

« As a boy I was a keen Isaac Asimov fan. In particular I read and reread his two most famous series: the *Foundation* trilogy and the *Robots* series. Oddly enough, the psychohistory of Hari Seldon, the chief protagonist in *Foundation*, was probably my earliest exposure to the idea of emergence in social systems. As Seldon puts it, although the behaviour of individuals is hopelessly complex and unpredictable, the behaviour of mobs, and even civilizations, is amenable to *analysis* and *prediction*. As fantastical as it was when conceived in the early 1950s, Asimov vision is remarkable in its anticipation of much that the study of complex systems attempts to do today », Duncan Watts, *Six Degrees: The Science of a Connected Age*, 2004.

En science, la théorie et l'expérimentation se polissent l'une l'autre, comme un lent et patient travail artisanal. Il s'agit d'y faire coïncider l'expérimentation et les hypothèses – ou les prévisions théoriques – qui donnent à l'objet de science sa forme et ses propriétés. Et souvent le chemin est long pour y parvenir, un chemin sinueux fait d'allers et de retours entre les *mots* – la théorie – et les *choses* sur lesquelles s'exerce sa vérification dans un dispositif expérimental. Il est rare en science que les données de l'expérimentation confirment *exactement*, et *d'un coup*, des prévisions théoriques ou des hypothèses. On y passe souvent son temps à aller de la théorie à l'expérimentation, à l'observation, à la modélisation – enfin à toutes ces opérations de production d'informations dirigées vers l'objet de recherche – afin d'affiner la démarche. Dans un sens, on adapte les hypothèses ou la théorie aux données expérimentales ; mais, dans l'autre, on modifie aussi le dispositif d'expérimentation pour le faire « coller » aux hypothèses. Ainsi, on peut considérer que la science est parfaitement *circulaire*, comme les côtés pile et face d'une même pièce. Cette circularité peut fasciner ou inquiéter, c'est selon. L'épistémologie et l'histoire des sciences ont pour vocation de l'interroger pour illustrer la relativité de la connaissance scientifique du moment,

établir la variation historique des théories et, surtout, considérer le savoir comme le produit d'un dispositif d'observation. C'est tout le regard englobant de l'historien, de l'érudit riche d'enseignements, fascinés tous les deux par la circularité de la science mise à distance comme objet de réflexion.

Personnellement, je me suis toujours un peu ennuyé dans cette tradition spéculative et érudite, celle pourtant d'où je viens à titre de chercheur en sciences humaines et sociales. Je préfère aujourd'hui partager l'enthousiasme (et la naïveté sûrement) de ceux qui cherchent volontairement à briser les équilibres dont nous avons hérité entre hypothèse théorique et dispositif expérimental, ce qu'avait justement mis longtemps à polir la tradition. Je partage avec d'autres cette sorte de pari : si l'effet de circularité est si fort dans la construction de l'objet scientifique, admettons qu'il est alors possible de le *subvertir*, non pas du côté pile de la théorie, mais du côté face du dispositif et des données de l'expérimentation. Cela fait maintenant longtemps qu'au contact des ingénieurs de l'information et des systèmes j'ai fait ce pari (insensé pour un chercheur) qu'une modification profonde de nos instruments d'observation associée à des données de nature radicalement inédite pouvait déboucher sur de *nouveaux objets* de connaissance et l'observation de *nouvelles propriétés* dans les données. C'est là tout l'enthousiasme des périodes d'invention, l'annonce de champs nouveaux à explorer et des instruments à inventer et adaptés à un nouveau type de données. Une époque de fascination pour les prouesses techniques des ingénieurs, et la promesse des théories à venir.

Et c'est exactement ce que vivent aujourd'hui les sciences, et les scientifiques. Avec l'arrivée des données numériques et l'avènement du Web depuis peu (du moins à l'échelle de l'Histoire), les innovations issues des technologies de l'information interrogent désormais de bien des façons l'activité scientifique, en particulier dans la nécessité de revoir certaines de ses bases expérimentales ou de ses instruments de prédilection (*in silico* désormais). Derrière le phénomène croissant des *data sciences*, se cachent d'innombrables pistes – plus ou moins chanceuses – d'innovations technologiques dont certaines modifient nos connaissances comme d'ailleurs les formes de nos sociétés. Rapidement, et peut-être *profondément*. Personnellement, j'aime cette période foisonnante de la « révolution numérique » où, au fond, les enseignants de nos universités n'en savent guère plus que leurs étudiants sur bien des sujets touchant aux nouvelles technologies, où la recherche scientifique commence par une forme d'ignorance partagée et où les ingénieurs (plus que les théoriciens) occupent le devant de la scène avec leur bricolage inventif et surprenant.

Si, de surcroît, on fait du Web, non pas un *vecteur* de la connaissance scientifique, mais un *objet* de recherche à part entière (les *web sciences*), on entre donc dans

un domaine encore plus incertain où tout doit être imaginé et construit, depuis le réseau, les *data*, leur extraction, leur stockage et leur traitement, les mesures et les algorithmes associés (métriques), leur enrichissement (en les croisant par exemple) sans oublier l'empilement local de logiciels ou d'applications accessibles en ligne où l'on retrouve... nos *data* de départ. Oui, je sais, je sais : on m'objectera qu'il existait des formes de réseaux *avant* le Web, que les méthodes utilisées comme la théorie des graphes existent *depuis* Euler au XVIII^e siècle et que les sociologues et les scientomètres avaient *déjà* conçu des méthodes et des algorithmes efficaces pour décrire le social sous forme de *sociogrammes*. En somme, si j'entends bien le point de vue, rien de nouveau ne serait apparu hormis le développement exponentiel des puissances de calcul et d'un réseau planétaire. Oui, c'est exactement cela, mais à une *différence majeure* : la quantité de données traitées aujourd'hui sur ce réseau distribué de *data* peut conduire à un *changement qualitatif* d'objet de recherche. Ceux qui participent aujourd'hui au développement de la science des réseaux (*network sciences*) en ont parfaitement compris le principe, notamment aux États-Unis et plus largement tous ceux qui se réclament des *data sciences*. Ils partagent cette idée qu'étudier le Web (et même le cartographier !), c'est peut-être faire émerger de nouveaux objets de connaissance, des hypothèses inédites. Le Web est aujourd'hui un immense terrain expérimental, en évolution rapide, qui représente un défi technique permanent. Sur ce point, il faut avouer que l'exploration du Web permise par les technologies numériques dédiées aux réseaux ne débouche pas encore sur la production de nouvelles théories, articulées et certaines. Mais l'effervescence est là, dans un champ foisonnant de nouvelles hypothèses de travail adossées à des instruments inédits et des *data* en masse, d'une grande variété et en évolution continue. Le plus frappant en cette période d'innovation scientifique et sociétale, c'est la vitesse à laquelle semblent se produire les mutations, comme si le temps s'accélérait soudainement dans l'invention de nouvelles formules d'association entre hypothèse et données, entre théorie et expérimentation, entre les mots et les choses désormais *numérisés*. Le temps du polissage viendra ensuite, l'énergie se concentrant d'abord sur la recherche d'une rencontre heureuse, et inédite, entre hypothèses (en cours d'élaboration), instruments (en cours de développement) et données (en cours d'extraction) dans l'esquisse des nouveaux domaines de recherche. Et il arrive que cette rencontre surgisse d'un coup, dans un alignement parfait de ces trois éléments.

Pendant les mois que j'ai passés à l'Institut de l'information scientifique et technique (INIST CNRS), je me suis lancé dans un travail long (et insensé !) d'identification

et de catalogage des sites web qui, à l'époque, donnaient accès à des stocks de données scientifiques (presque) brutes. Plongé dans ce temple de la rigueur et de l'exactitude qu'est l'INIST où officiaient documentalistes et ingénieurs de recherche, je me suis initié aux pratiques et aux méthodes de la bibliométrie et ses structures traditionnelles de graphes, notamment appliquées aux mesures de citations entre chercheurs, entre revues ou entre laboratoires. Mais j'avais souvent à l'esprit ce projet de recensement qui, je l'espérais, pourrait peut-être déboucher sur quelques trouvailles, surtout cartographiques. Nous avons désigné ce projet sous le nom de Web Datarium. Pendant plusieurs mois, j'ai repéré et parcouru plus de 3 500 sites web scientifiques, pour en retenir, au final, plus de 700.

Le modèle que j'avais en tête consistait à parcourir le Web manuellement en essayant de me concentrer sur les liens qui menaient de site web scientifique à site web scientifique. Je devais me focaliser sur ce seul domaine, avec l'espoir de voir apparaître un paysage continu et où l'on pourrait naviguer en tous sens, sans rupture. Quand cela le nécessitait, je n'hésitais pas non plus à recourir à Google pour multiplier mes points d'entrée sur le réseau afin de vérifier que toutes les URL pertinentes que je trouvais étaient bien reliées entre elles. J'étudiais donc à la fois les contenus des sites et leurs liens hypertextes. Pour les plus jeunes, disons que ce Web scientifique documentaire était du type 1.0 : des sites statiques et des pages fixes, rarement une animation et une structure type qui permettaient d'accéder, depuis la page d'accueil, aux pages recherchées où figuraient des accès à des bases de données en ligne. Parfois, il m'arrivait même de trouver encore des pages web sur fond gris, écriture noire, lien en bleu. On aurait dit une couche lente et rudimentaire du Web où s'établissaient les liens primordiaux d'un réseau de connaissances. Quelque chose comme une couche basse dans laquelle le Web plante ses racines à titre de système d'information, ce qu'annonçait notre modèle du Web en couches. À l'époque, les sites n'étaient pas reliés à des blogs (comme aujourd'hui avec les témoignages des scientifiques) ou à Twitter pour signaler la diffusion d'un travail. J'ai longtemps navigué dans cet univers fascinant des données produites par des capteurs en mer, des radars, des sondes, des mesures de température ou de pression en passant par les images satellites, les données astronomiques ou celles (en temps réel) des incendies de forêt dans le monde. J'intégrais aussi à mon corpus les accès aux bases de classement des espèces, aux catalogues de leurs gènes sans oublier les molécules, les pathologies ou encore les matériaux et les médicaments. Les données, en tant que telles, ne constituaient pas mon objectif : je vérifiais seulement qu'elles étaient *accessibles en ligne* en prenant soin d'identifier dans une liste l'URL de la page d'accueil du site ainsi que l'URL de la page donnant accès à ces fameuses informations dont le foisonnement m'étonnait. Elles étaient à la fois si nombreuses et si étranges pour moi que j'en vins à l'idée qu'il ne pouvait pas y avoir un ordre, une logique, une organisation générale

qui pouvait gouverner un tel déploiement en ligne. Peu importe, après tout, puisqu'il suffisait dans les objectifs du projet d'en dresser la liste la plus exhaustive possible, si possible au format Excel, et de mettre en forme logiquement les résultats avec une arborescence. En somme, des listes raisonnées ou des tableaux de classification, qui ne rendraient jamais compte des formes de connexion qui animaient en ligne toutes ces ressources. Durant cet arpentage fastidieux du Web des données scientifiques, mon univers de travail se trouva rapidement enrichi de nouveaux instruments. Mathieu Jacomy venait d'entreprendre un long travail de mise en forme des méthodes associées à son outil, le Navicrawler. Techniquement, il s'agissait « juste » de l'adjonction à la version courante de Firefox d'une série de *plugins* d'enrichissement des parcours de navigation : repérage et stockage automatique des liens hypertextes identifiés sur les pages web visitées, construction d'un graphe du parcours sur le Web, possibilités de soumettre au Navicrawler une liste préétablie d'URL pour extraire la structure hyperliée associée... Un instrument privilégié pour tous ceux qui veillent et observent.

D'un autre côté, Mathieu Bastian et Sébastien Heymann, qui m'avaient accompagné à l'INIST à la fin de leurs études d'ingénieurs, étaient au développement de la dernière version de Gephi. Parmi toutes les améliorations apportées à l'application, l'une fut pour moi décisive : l'algorithme de spatialisation Force Atlas 2¹. L'algorithme Force Atlas 2, comme son prédécesseur le Force Atlas, est un algorithme de type *force directed*, autrement dit un type de spatialisation d'une structure de graphe qui fonctionne en exploitant la distribution des liens entre des éléments pour les transformer en force d'attraction et de répulsion. Avec Gephi, comme dans d'autres applications d'analyse de graphe, les algorithmes de ce type figurent en bonne place aujourd'hui dans la boîte à outils des instruments de spatialisation.

1. M. Jacomy, T. Venturini, S. Heymann et M. Bastian, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software", <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679>.

Image 7.2

Un même réseau visualisé au moyen de différents algorithmes de spatialisation.

L'algorithme présentait plusieurs qualités qui m'apparaissaient comme les clefs de voûte d'une possible cartographie du Web, depuis de petits corpus comme on aime à les construire qualitativement en sciences sociales jusqu'aux masses d'URL archivées dans les moteurs de recherche. Force Atlas construit l'espace de projection selon des principes qui m'ont semblé correspondre, dès le départ, à l'instrument dont j'avais besoin : il fait apparaître rapidement les effets de concentration et de périphérie, il est particulièrement sensible aux effets de *clusters* (groupements locaux d'éléments) et, plus important encore, son développement s'est nourri de façon continue de retours d'expériences. Il s'agit donc d'un instrument dédié à l'accompagnement de l'interprétation, profilé pour concentrer l'analyse sur les effets de *distance* ou de *proximité structurales*. Il ne s'agit pas d'une projection euclidienne, mais bien d'un vaste système de positionnement relatif des éléments les uns par rapport aux autres. D'où quelques-unes des caractéristiques qui font le succès de cet algorithme. Il n'est pas conçu, en premier lieu, comme une « boîte noire » sans possibilité d'intervenir sur le processus et les paramètres de ce jeu de force : au contraire, il est presque intégralement paramétrable manuellement depuis la console Gephi tels sa vitesse d'exécution, le réglage des forces de répulsion et d'attraction, son échelle, sans oublier quelques améliorations décisives comme celle d'Andreas Noack avec son LinLog energy model qui épouse les principes de méthodes statistiques de détection des communautés de type Modularity. Les *patterns* générés par l'algorithme peuvent être enrichis dans l'environnement Gephi pour y manifester différentes métriques comme les mesures de hiérarchie des éléments (degré ou

nombre de liens, centralité), la distribution des couleurs et des familles de nœuds (justement avec l’algorithme Modularity) ou encore la possibilité d’isoler une partie du graphe (notamment un *cluster*) pour le projeter dans un nouvel espace de travail.

Exemples d’étapes de travail lors de l’exploration visuelle d’un réseau.

Le prix à payer pour cette souplesse est évident : l’algorithme Force Atlas est particulièrement gourmand en temps de calcul. En d’autres termes, il est lent. Le déploiement de forces d’attraction et de répulsion, surtout à grande échelle, implique des calculs importants de relativité de positionnement de tous les éléments d’un fichier de graphe. Au final, le système finit par trouver un état de stabilité relative avec cette caractéristique que le processus est continu : c’est l’*effet élastique* du Force Atlas 2.

Pendant que l'algorithme continue de calculer les positions, on peut « tirer » sur un nœud qui reprend rapidement sa place d'origine (nous avons créé un fichier de graphe en forme de tamis régulier pour illustrer le principe) :

Image 7.4

Poster du projet Web Datarium.

Il ne s'agit pas d'une facétie d'ingénieur : un algorithme qui fonctionne de façon continue se révèle particulièrement efficace pour spatialiser des données sous forme de flux (si l'on pense, par exemple, à Twitter et à son API). Dès que ces instruments furent testables, j'élaborai immédiatement une petite chaîne locale de traitement de l'information : j'envoyais vers le Navicrawler mes 750 URL de sites web scientifiques dont il releva les liens externes. J'avais donc en main une structure de graphe et je ne conservais que les liens entre URL présentes dans mon corpus (les liens menant vers des sites externes ont été éliminés, mais temporairement seulement). Le fichier de graphe fut ensuite analysé avec Gephi et spatialisé avec Force Atlas 2. Le résultat ressemblait à ce que j'attendais : un *cas exemplaire*.

Image 7-5

Poster du projet Web Datarium.

Pour qui est à la recherche d’une forme du Web, qui plus est cartographiable ou analysable avec des algorithmes comme HITS, l’univers affiché offrait un terrain remarquable d’efficacité technique. Tout nous est apparu impeccablement rangé, plié à nos instruments et à nos méthodes et répondant à nos hypothèses de prédilection sur le principe d’agrégation de documents.

Les URL d’accès aux bases de données scientifiques s’agrégeaient en localités thématiques aisément identifiables. Comme j’avais pris soin dans mon fichier Excel d’origine de noter dans une colonne spécifique quelques mots-clés indiquant le champ scientifique majeur des sources retenues, il fut facile d’associer chaque localité à une sorte de toponyme. Mes *localités thématiques* étaient identifiables sur la carte par une batterie de mots-clés, informations que ne délivraient ni les liens ni les nœuds (à moins, évidemment, de fréquenter régulièrement les URL). Au final, le procédé permettait d’obtenir une sorte de classification ascendante des grands domaines de recherche, la logique des liens hypertextes permettant de hiérarchiser, comme de grouper les URL.

Image 7.6

Détail de Web Datarium

Au centre de la carte trônait le site de la National Science Foundation (NSF). Une excellente illustration de la *domination* américaine sur l'univers des sciences en ligne, et sûrement en dehors aussi. Enfin, le terme n'est peut-être pas le bon : si quelques liens en sortaient pour pointer vers des organismes gouvernementaux (.gov) qui servaient de *hubs* locaux (un pour les sciences de la terre, un pour la génétique, un pour la biologie, un pour l'environnement...), l'essentiel de la connectivité qui reliait le site de la NSF au réseau était généré par des liens *entrants*. À ce niveau-là du Web, disons dans cette grande région de la science, le site faisait figure de pure *authority* dont la position se renforçait probablement au fur et à mesure que se développaient des projets fédératifs en ligne. Le terme d'*autorité*, plus que *domination*, inclut l'idée d'un processus de reconnaissance partagée, produit spontanément par une foule d'acteurs. Exactement ce que j'ai fait en ouvrant mon blog en pointant vers mes auteurs favoris. Dans l'univers du partage des données scientifiques en ligne (aujourd'hui une nécessité dans la plupart des domaines), il semble que la moindre

initiative « locale » ou « nationale » finit par faire rayonner un peu plus les grandes agences gouvernementales américaines dont la NSF est la figure de proue. Par exemple, dans la partie haute de la carte et sur la droite, le NIH (National Institute of Health) semble régner sur les sciences du vivant tandis que dans la partie gauche et en bas de la carte se dessinent les terres de la NASA, de ses instruments d'observation et de son ingénierie.

Image 7.7

Détail de Web Datarium.

On doit probablement cette visibilité (nombre de liens) et cette centralité (des passages obligés en matière de navigation hypertexte) des sites en .gov de la recherche américaine à une combinaison parfaitement maîtrisée d'ancienneté (avec les premières politiques de libération des données aux États-Unis), de dynamique de communication sur le Web, et d'ingénierie des *data* et des systèmes d'information conçus dès le départ comme distribués, contributifs et délocalisés (avec sa pléiade d'applications côté clients et non plus localisées dans un serveur central). Il y a là toute une organisation de la cyberinfrastructure qui ne doit rien au hasard et révèle un sens (américain) inné de l'occupation des territoires numériques. Un système en ligne comme EOSDIS (Earth Observing System Data and Information System) fut imaginé dès 1989 ! À l'époque, la France sortait à peine de son rêve de Minitel et l'Europe cherchait encore l'organisation de sa recherche...

Au-delà, ou en deçà de cette colonne vertébrale NSF-NIH-NASA, l'univers des agrégats ou des localités thématiques semble dérouler en toute logique des complémentarités remarquables. Toute la logistique de la cyberinfrastructure qui permet le travail scientifique en ligne (les « *cyberbricks* » : les données aux formats divers et à différents niveaux de structuration, dynamiques et à différentes échelles, données relationnelles, taxinomies, formules, outils de calcul et d'exploration, moteurs de termes...) se décline de façon circulaire, comme les heures de l'horloge. À midi, la génomique et la biologie, à 3 heures les océans et leur biodiversité. Donc, entre

les deux (disons vers 1 h 30), la biologie et la biochimie marine avec ses variables environnementales (comme la chaleur des océans ou les aléas climatiques). À 6 heures, les sciences géographiques et leur ingénierie des mesures et de l'observation. Aux alentours de 6 h 30, la climatologie, la paléoclimatologie et la problématique du *climate change*, une question historiquement scientifique avant d'être politique. Un peu avant 9 heures, les sciences et les techniques d'observation de la terre comme système géophysique et la dynamique des écosphères. À neuf heures passées, les mêmes questions débouchent cette fois à l'échelle de l'univers (l'origine de l'univers, l'astrophysique, l'exobiologie...). Enfin, la question de l'origine de la vie vient achever la boucle en débouchant sur l'infiniment petit, les gènes et les cellules.

On aurait dit que la carte contenait à la fois le dénombrement de grands domaines scientifiques, mais aussi leur combinatoire. Cette vue apparaissait presque totalisante : produite par le jeu des URL et des liens hypertextes et mise en scène avec Force Atlas 2, elle nous rappela ce fameux théâtre de mémoire imaginé par Giordano Bruno à la fin de la Renaissance² ou, encore, ces roues de la connaissance du Moyen Âge censées contenir toutes les formes de savoirs et leurs combinaisons possibles³. Parti d'un laborieux travail de recensement et de qualification, voici que nous étions projetés aux frontières de l'histoire, de la philosophie des connaissances ou de l'ésotérisme, tout d'un coup, une fois les données passées au filtre de l'algorithme Force Atlas 2.

Image 7.8

Détail de Web Datarium.

2. F. A. Yates, *L'art de la mémoire*, Paris, Gallimard, 1975.
3. M. Carruthers, *Machina memorialis. Médiation, rhétorique et fabrication des images au Moyen Âge*, Paris, Gallimard, 2002.

Ce qui nous fascinait à l'époque, plus que la forme elle-même de la carte, c'était de contempler toutes ces URL et leurs contenus : leur nombre et leur diversité démontraient la façon dont notre univers était systématiquement sondé, mesuré, observé, analysé, depuis les molécules jusqu'aux astres lointains. On aurait dit une nouvelle étape historique dans la quête de recensement généralisé des êtres et des choses, l'« âge classique » de Michel Foucault, avec ses listes taxonomiques, les arborescences où viendront ensuite se loger les pyramides hiérarchisées (totales) des êtres vivants. L'horizon d'observabilité des phénomènes nous paraissait immense à naviguer dans cet univers de bases de données ou de connaissances en réseau. Avec quelques étapes dont je me souviens, par exemple la possibilité de suivre des incendies de forêt en temps réel ou encore le *crowdsourcing* scientifique avec l'anecdote de Hanny Van Arkel, un professeur des écoles en Hollande, qui a trouvé avec ses élèves une galaxie via le Sloan Digital Sky Survey et qui porte aujourd'hui son nom.

Le Web et l'Internet existaient donc comme *scientific backbone*. Notre photographie de 2008 annonçait le développement actuel des *e-sciences* dont parlait Jim Gray à l'époque⁴. Il voyait, dans cette infrastructure en réseau et ses technologies, l'équivalent pour la science d'un nouveau paradigme, le quatrième après la science expérimentale, la science théorique des modèles, la science de la simulation et des technologies numériques. L'idée qu'il développait consistait à imaginer comment naviguer dans cet espace accompagné d'instruments d'analyse portatifs (des *web services* dirait-on aujourd'hui) alors que les *data* resteraient là où elles sont dans les bases accessibles. Ce que nous avons photographié à l'époque n'était que l'ossature primaire (des bases reliées entre elles via des liens hypertextes statiques) où devaient s'hybrider plus tard les *data* à peine captées et les connaissances en tant que telles : «If you have a schema for things, you can index the data, you can aggregate the data, you can use parallel search on the data, you can have ad-hoc queries on the data, and it is much easier to build some generic visualization tools» précisait ainsi Jim Gray.

Notre carte ne faisait donc que dévoiler la forme fixe et primaire du réseau de connexions entre différents silos de données. Vu l'ampleur et la diversité des types de données, on aurait dit qu'il y avait à l'œuvre sur le Web une entreprise généralisée de recensement, dont la métrologie serait la science générique. L'homme, la nature, l'univers : rien ne semblait échapper aux capteurs, aux données, aux mesures, à cette foule de projets de recensement (dont certains étaient contributifs comme l'identification des espèces de biotopes locaux où les internautes étaient appelés à se mobiliser). C'est sur ce tamis, presque fixe, des domaines scientifiques que viennent

4. T. Hey, S. Tansley et K. Tolle, "The Fourth Paradigm: Data-Intensive Scientific Discovery", Microsoft Corporation, 2009, <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>.

aujourd'hui se brancher les outils de calculs embarqués qui vont chercher les données. Une foule de processus plus dynamiques peuvent désormais s'y déployer : à partir de formats relativement communs (et donc des « données interopérables »), des requêtes dynamiques peuvent entrer en jeu via les API (les clefs d'entrée de chacun des silos informationnels), des applications locales depuis un simple ordinateur ou des *plugins* avancés depuis nos navigateurs peuvent se déplacer dans cet univers et en exploiter les richesses. Le foisonnement de tous ces *workflows* scientifiques sur le réseau serait passionnant à cartographier.

Nous n'avons pas poussé plus loin notre expédition dans cet univers fascinant des données et des connaissances scientifiques. Là aussi, le travail de cartographie aurait dû conduire au développement d'un moteur de recherche dédié où aurait pu figurer une carte interactive de synthèse et de rassemblement. Elle aurait révélé la forme ou l'organisation (presque rigide dans sa logique de déclinaison des domaines scientifiques) qu'avait contribué à produire l'activité collective de milliers de chercheurs à travers le monde. De retour à nos occupations quotidiennes (chacun devant justifier son emploi dans des missions qui, pour tout dire, n'étaient guère passionnantes), nous étions cependant satisfaits de cet alignement presque parfait des processus (analyse des URL, extraction de leurs liens réciproques via un *crawler*, calculs en base sur nos archives, projection sous Gephi avec le Force Atlas 2) qui ouvrait sur un espace jusque-là jamais cartographié et dont l'organisation nous avait frappés. Et, dès les premières projections de cette géographie des connaissances scientifiques sur le Web, nous avons laissé libre cours à nos pensées sur la forme et le rôle des instruments de synthèse des savoirs, sous forme de roues, de cartes ou de théâtres, ou prospectivement en découvrant ce qu'annonçait Jim Gray sur l'avènement des *e-sciences*. C'est dans ces moments-là que l'on mesure le pouvoir heuristique de la cartographie d'informations, lorsqu'elle disparaît à titre de filtre ou d'instrument. Notre alignement de processus de traitement des *data* ne fonctionnait donc qu'en devenant transparent. *Invisible*.

Chapitre 8. *Broadcast* ou viralité ?

« Whenever nonlinear elements are hooked together in gigantic webs, the wiring diagram has to matter. It's a basic principle: Structure always affects function. The structure of social networks affects the spread of information and disease; the structure of the power grid affects the stability of power transmission. The same must be true for species in an ecosystem, companies in the global marketplace, cascades of enzyme reactions in living cells. The layout of the web must profoundly shape its dynamics », Steven Strogatz, *Sync: The Emerging Science of Spontaneous Order*, 2003.

Lorsque Mathieu Bastian, le *lead developer* de Gephi, reçut le Duke Award en 2010 au JavaOne, j'étais ravi de voir reconnu au niveau mondial tout le travail des jeunes ingénieurs et des membres du consortium¹ créé à l'époque pour soutenir le développement de l'application. Ce premier prix dans la catégorie « Innovative Technical Data Visualization » venait récompenser nos efforts et quelques trouvailles en matière de visualisation de graphes. Je ne me lasse pas de revoir sur Vimeo² les quatre minutes allouées à la prestation remarquable de Mathieu : il avait choisi de montrer quelques-unes des fonctionnalités clés de Gephi en prenant pour exemple le code Java lui-même (tous les *packages* du langage reliés entre eux) et la présentation de la *timeline* vint couronner l'exposé. Il laissa tourner l'algorithme Force Atlas 2 en continu et fit entrer dans le jeu de spatialisation l'ensemble des huit versions successives de Java de l'époque. Visuellement, l'effet était remarquable : les uns après les autres, des dizaines de nœuds vinrent se placer dans l'espace en cherchant leur position optimale, mus par le jeu de plus en plus complexe des forces d'attraction et de répulsion qu'imposait le Force Atlas 2 aux données. Enfin, le temps dans les graphes semblait maîtrisé ! Du moins, il était maintenant *manipulable* depuis une ligne en bas de l'écran qui fonctionnait comme le cinéma, et avec la même magie.

Je me pris à rêver de pouvoir disposer d'archives exhaustives et de la puissance de calcul nécessaire pour admirer l'évolution globale du Web comme système depuis le *big bang* de départ, la première page HTML de Tim Berners-Lee³. On y verrait l'expansion du Web comme le renforcement de ses forces locales de gravité, la conservation d'un diamètre moyen faible pour le nombre de documents présents, l'existence continue d'un *giant component* où tout est (presque) relié à tout ce qui définit le Web

1. <https://consortium.gephi.org/>.

2. La vidéo n'est plus en ligne actuellement.

3. <http://info.cern.ch/hypertext/WWW/TheProject.html>.

comme un *espace continu* (ce qui rend possible l'existence de moteurs de recherche généraliste comme Google) jusqu'aux stratégies locales adoptées par les nouveaux entrants dans le système comme je l'ai fait avec mon blog personnel. On y verrait aussi, en toute probabilité, la spécialisation progressive de certains nœuds du système, comme ceux qui figurent dans sa couche haute avec les différents types de réseaux sociaux grands pourvoyeurs de connectivité globale et les évolutions particulières de certains agrégats (en pensant aux ressources web diffusées par les adeptes de l'Église de scientologie). Le Web serait-il sujet à une évolution progressive ou bien à une série de phases ou de crises intenses ? Quelles sont les lois qui en règlent le développement ? Pour une fois, l'évolution du système ne serait pas déduite de lois statistiques, mais d'une forme de rétrospective dynamique et visuelle, et contrôlable depuis la console cinématographique d'une *timeline*.

Malheureusement, il n'existe pas d'archives de ce type (à ma connaissance), ni même de programme scientifique pour traiter de questions aussi « métaphysiques ». La *timeline* de Gephi ne fonctionnera donc que sur des jeux de données relativement restreints, et seulement s'ils contiennent des informations temporelles. Pour ceux qui ne sont pas ingénieurs, la *timeline* n'est qu'une fonction de plus dans Gephi. Pour ceux qui travaillaient sur l'application à l'époque, ce fut un redoutable défi technique. Lorsque je l'ai vue fonctionner pour la première fois, la *timeline* avait quelque chose de fascinant, car, avec les deux dimensions du plan ou de l'interface, elle semblait achever de nous donner le pouvoir sur l'ensemble des variables de l'univers des *data*. Elle a quelque chose à voir avec le cinéma et avec ce point de vue omniscient comme l'a construit la technique de la perspective en peinture. Les *data* ont trouvé avec elle le dernier élément de leur mise en scène, avec des solutions parfois magnifiques dans le domaine de l'interface et de la *dataviz* sous forme de curseurs, de sphères ou de points assurant le contrôle du « temps » dans les jeux de données.

Mais le temps sur le Web reste encore une question si complexe qu'un livre ne suffirait pas à en examiner tous les aspects⁴. Une fois passé ce moment de fascination pour les pouvoirs de la *timeline*, je me rendis compte à quel point nous ne faisons qu'effleurer la question du temps dans les graphes avec son cortège de problèmes théoriques et technologiques. Un point en particulier me semblait central pour comprendre les différentes dimensions du temps sur le Web (et donc pour imaginer des instruments de contrôle de la temporalité efficaces dans le cadre de notre projet de recherche technologique). La vue dynamique présentée par Mathieu Bastian lors du JavaOne permettait d'apercevoir l'évolution de la structure en termes d'apparition/disparition de nœuds et de liens (ce que l'on appelle aussi *information IN the net*

4. M. Newman, A.-L. Barabási et D. Watts, *The Structure and Dynamics of Networks*, Princeton (NJ), Princeton University Press, 2006.

en désignant par-là la structure elle-même du réseau), mais ne pouvait pas rendre compte de la façon dont peut circuler une information sur ce réseau (ce que l'on désigne par *information ON the net*). Autrement dit, la *timeline* de Gephi portant sur les évolutions de structure devait être clairement considérée comme un premier pas dans la quête d'une représentation plus exhaustive du temps et de ses différentes dimensions, et non comme une finalité. La cartographie de cette sorte de « temps interne » ou de la dynamique de la circulation de l'information devint donc un second objectif. Ce qu'il reste toujours aujourd'hui.

À l'examen, ces questions émergent dès que l'on utilise des graphes qui *produisent d'eux-mêmes* une forme de temps qui n'est pas réductible à la vue cinétique et extérieure de la structure en évolution. Je ne sais pas ce que vaut la distinction qui suit, mais elle me semble éclairer certaines particularités des graphes appliqués aux réseaux comme le Web : ces graphes ouvrent un espace particulier dans lequel le calcul des chemins entre les nœuds à travers leurs liens, donc des *distances* et des différentes *propriétés* d'une *topologie*, est aussi un calcul du *temps*. La distance topologique entre A et B, c'est aussi le temps d'aller de A vers B (sans parler du retour de B vers A qui rend alors l'espace *réversible*, une figure essentielle des liens hypertextes). La structure dessinée par l'ensemble des nœuds et des liens d'un graphe (de pages web, mais pas seulement) peut donc être considérée sous l'angle des *chemins*, donc des *déplacements*. Nous sommes alors dans une logique de flux. Je peux aussi élargir le principe et considérer le lien de A vers B comme une transaction commerciale (par exemple en pondérant les liens) ou comme l'expression d'un système d'interactions sociales. L'image de la circulation de produits financiers « nocifs » dans le réseau des transactions permises par une infrastructure mondialisée me semble correspondre au principe.

Image 8.1

Steven Strogatz a fait de la synchronisation des éléments d'un système complexe son champ de recherche. Ici, il présente les principes de synchronisation d'une foule dans un stade de football américain dans le documentaire *The power of Six Degrees: How Kevin Bacon cured cancer*. Il a travaillé avec Duncan Watts sur différents modèles de graphes, notamment celui en *power-law* dont les propriétés topologiques (diamètre court, clusterisation des composants en agrégats) semblent expliquer la capacité de certains systèmes complexes – comme le Web – à vivre des périodes de synchronisation. « At the heart of the universe is a steady, insistent beat: the sound of cycles in synchrony » (S. Strogatz, *Sync: The Emerging Science of Spontaneous Order*, 2003).

Dans tous ces cas de figure, la topologie du réseau semble *contenir* une forme de temps traduite des distances, et donc des propriétés statistiques connues dans le cas du Web. Par exemple, quand je calcule des mesures de *diamètre moyen* d'un réseau, ou d'autres propriétés comme sa *densité* (nombre réel de liens / nombre possible de liens) ou des degrés de *centralité* pour chacune de ses composantes, les liens constituent mon unité de mesure de l'espace, mais aussi du temps du déplacement entre les nœuds. Le temps est d'ailleurs potentiellement inscrit dans certaines figures statistiques déjà rencontrées, par exemple avec la mesure du caractère d'*authority* d'une page web avec HITS qui, comme le souligne Kleinberg, tend à se renforcer au cours du temps. À vrai dire, la plupart des travaux scientifiques en *web mining* contiennent une réflexion potentielle sur l'évolution temporelle du système, en particulier quand il s'agit de questionner la conservation de lois statistiques calculées pour l'essentiel sur des jeux de données statiques⁵.

Dans cette perspective, il faut considérer que cette dynamique des flux est contrainte, presque *produite* par la topologie. Envisagé sous l'angle des flux, comme système transactionnel ou comme un immense terrain d'interaction sociale (via des documents ou des informations), le réseau retrouve une dimension dynamique que les graphes « à plat », et la cartographie, ont peut-être trop rapidement figée dans une forme statique. Ainsi, dans l'étude de processus distribués comme les flux physiques, les charges électriques, les transactions commerciales, des contaminations virales, le lien entre A et B peut figurer l'*événement-réseau élémentaire* dont A et B sont les

5. J. Leskovec, J. Kleinberg et C. Faloutsos, "Graph Evolution: Densification and Shrinking Diameters", *ACM Transactions on Knowledge Discovery from Data*, 1 (1), 2007, <https://arxiv.org/abs/physics/0603229>.

bornes. Appliqué à grande échelle, le principe ouvre des perspectives fascinantes en matière de gestion des transports, de l'énergie ou d'épidémiologie. Sous l'angle de ce type de dynamique, il faut désormais considérer le Web comme on le fait en biologie à propos du cerveau humain : une structure topologique de neurones et de connexions, mais aussi la dynamique de l'activation de certaines configurations neuronales via des messages chimiques ou électriques. De nombreux projets de cartographie interactive existent aujourd'hui fondés sur l'étude des flux et de la circulation de l'information et qui ont pour vertu (affichée ou non) de révéler la morphologie globale des réseaux (notamment des réseaux sociaux de tous types)⁶.

Image 8.2

Interface de visualisation des échanges épistolaires entre philosophes tout au long du XVIII^e siècle et produite dans le cadre du projet Mapping the Republic of Letters du département Digital Humanities à Stanford. Chaque point lumineux représente une lettre qui circule le long des lignes qui relient les adresses postales où séjournent les philosophes en Europe. Le Siècle des lumières peut aussi s'envisager comme un réseau distribué d'acteurs qui s'influencent mutuellement à une époque où l'activité épistolaire prenait une place majeure dans la distribution des idées. Dans ce projet, c'est la circulation dynamique des lettres qui révèle la topologie du réseau (Giovanni Cesarini, Dan Edelstein, Paula Findlen, Caroline Winterer, 2009).

Cette quête est aujourd'hui à peine entamée en raison des verrous techniques et théoriques qui l'accompagnent. En particulier, l'un des problèmes réside dans les méthodes et les instruments que nous pouvons déployer pour photographier les traces de cette dynamique sous forme de configurations topologiques actives, tout comme les *assemblées de neurones* dans le déclenchement d'une tâche cognitive. J'aime à considérer ce type d'objectif comme une *quête des empreintes* dont la collection permettrait d'évaluer en quoi un phénomène dynamique éclaire tout ou partie d'un graphe et selon quel(s) type(s) de propagation à travers une structure qui lui imprime ses propres contraintes. En d'autres termes, que nous apprend la structure sur le temps et, simultanément, que nous apprennent les phénomènes de propagation sur la structure ?

6. <https://www.youtube.com/watch?v=nw0oS-AOipe>.

Ce que j'appelle des *empreintes topodynamiques* livrera des informations précieuses sur les durées différentes des phénomènes observés, de la plus courte (A- > B distance 1 ou intervalle temporel 1) à la plus vaste en calculant tous les chemins possibles dans un graphe (par exemple, si l'on imagine un ensemble de nœuds contaminés de proche en proche par un virus, étape après étape). En d'autres termes, les nœuds d'un graphe s'allument-ils tous d'un seul coup ou bien par vagues successives, jusqu'à contaminer l'ensemble de la structure ? Et d'où vient le signal de départ, quel(s) type(s) de nœuds entame-t-il le processus ? Et au-delà, se posent toutes les questions (déjà connues) du type de réseau et de nœuds auquel on a affaire : les éléments d'une structure en réseau (du simple relais électrique dans un réseau de distribution de l'énergie à l'individu engagé sur les réseaux sociaux) ne sont pas passifs et une alchimie complexe se joue en chacun d'eux pour transmettre à leurs voisins une information, la transformer, voire la retenir sans la diffuser. Ces empreintes pourraient ainsi révéler l'*amplitude* ou la « surface » des phénomènes dynamiques si l'on considère le nombre de nœuds et de liens impliqués dans un phénomène dynamique en rapport avec le nombre total de liens et de nœuds du graphe. Un événement peut ainsi concerner seulement quelques nœuds locaux d'un agrégat ou bien l'ensemble du système. Pour comprendre la dynamique de l'information sur le Web, il faudrait donc pouvoir adosser des scénarios de propagation aux propriétés de structure, une clef méthodologique et théorique.

À ma connaissance, il n'existe pas encore de bibliothèque accessible de cas argumentés sur ce genre de phénomènes (notamment des chemins parcourus par une information sur le Web ou les réseaux sociaux), qui pourraient faire émerger des modèles d'empreintes topodynamiques ou valider des modèles théoriques à travers des jeux de données nombreux et divers. Pour être franc, je ne possède pas encore ce genre d'appareil photographique puissant qui m'ouvrirait sur des archives suffisamment riches et diverses pour établir des modèles de circulation ou de propagation de l'information sur le Web.

Mais des pistes d'expérimentations ont été ouvertes ces dernières années, notamment sous l'angle de l'étude des phénomènes sociaux propagatoires que l'on rapproche souvent de la notion de *viralité* ou d'*epidemics*. Dans ce champ des *web studies*, on considère le Web comme un *média social* dont s'emparent les internautes qui le transforment en espace d'interactions multiples à travers la production et la diffusion d'informations. Parmi tous les jeux d'interaction possibles, celui de la

production ou de la recherche *d'influence* a particulièrement retenu l'attention des chercheurs en sciences sociales, notamment les sociologues et les économistes qui se sont plongés dans l'étude de la diffusion de l'innovation, des opinions et des croyances, des communautés sociales sous l'angle du partage des valeurs idéologiques. Je n'oublie pas non plus le cortège d'experts en marketing, en stratégies d'influence, en communication publique, en *community management* ou en influence des marques qui ont l'œil rivé sur ce champ de recherche depuis l'avènement du Web 2.0. Et on comprend cet intérêt : une étude exhaustive des mécanismes de la diffusion d'une idée ou d'une information éclairerait la nature ou la forme de la structure sociale, mais constituerait aussi le premier pas sur la possibilité de l'orienter ou de la contrôler. Cependant, nous en sommes encore loin et il faut accorder un crédit très relatif à ceux qui font commerce de leurs capacités à « agir » ou à « prévenir » sur des phénomènes aussi dynamiques que la rumeur ou la réputation, des phénomènes qui restent encore mystérieux à bien des égards.

Si les interrogations théoriques demeurent aujourd'hui, les *data* arrivent. L'étude du Web considéré comme un espace d'interaction sociale a permis depuis quelques années le développement d'une infrastructure de recherche. C'est le cas notamment d'une foule de *start-up* dans le monde qui produisent jour après jour de grands corpus de *data* selon leurs domaines d'activité, là où les laboratoires publics de recherche ne peuvent rivaliser. De cet univers de l'observation du Web comme système social (notamment par les industriels de l'information) commencent à émerger au plan théorique et méthodologique des CSS dont se réclame aujourd'hui Duncan Watts notamment⁷. Comme le précise l'auteur dans ses plus récentes conférences⁸, le coût, l'échelle et la rapidité des expérimentations à partir des données d'usage ou les traces générées sur le Web par les activités en ligne renouvellent profondément les formes de l'expérimentation en sciences sociales. Ces données, et au-delà le *big data*, permettent de balayer aussi bien les niveaux *macro* que *micro* d'un phénomène et comportent souvent des informations temporelles. Une aubaine pour un *data scientist*, du moins si vous travaillez dans l'univers des *start-up*.

Ces avancées en termes d'ingénierie de captation ont démultiplié nos capacités d'observation, mais, pour le moment, sans avancée notable en termes de modèles théoriques, voire de prédiction. Duncan Watts rappelle d'ailleurs la faiblesse théorique de ce champ émergent des CSS⁹. Pour le dire autrement, nous n'en sommes qu'au

7. D. Watts, "Computational Social Science: Exciting Progress and Future Directions", *The Bridge on Frontiers of Engineering*, 43 (4), 2013, <https://www.nae.edu/106118/Computational-Social-Science-Exciting-Progress-and-Future-Directions>.

8. D. Watts, "Computational Social Science: Exciting Progress and Grand Challenges", https://www.youtube.com/watch?v=KgAi_LZOsx0.

9. <https://iriss.stanford.edu/css>.

début de la (longue) quête des empreintes topodynamiques qui pourraient alimenter en retour un ou plusieurs modèles théoriques. Les chercheurs de ce domaine et des équipes de R&D des *start-up* spécialisées dans l'étude du Web comme média social concentrent aujourd'hui leurs efforts sur une question centrale : quelle(s) forme(s) structurale(s) peut revêtir la diffusion d'une opinion, la propagation d'un virus informatique, l'adoption d'une innovation ? S'agit-il de phénomènes qui supposent une propagation comme on l'imagine pour des virus dans la société humaine ou bien d'un changement d'état rapide et instantané ? Modèle de diffusion de l'information de type *broadcast* ou modèle de type *viral* ? À l'évidence, il s'agit d'une question massive, mais centrale, notamment pour tous ceux qui considèrent les mécanismes de l'influence comme l'un des moteurs du Web sous l'angle de sa dynamique. À certains égards, c'est une évidence : le Web joue un rôle manifeste et puissant dans la propagation des idées (nouvelles ou anciennes), des mobilisations politiques (comme dans le cas des partisans du « non » au projet de Constitution européenne en 2005) ou des campagnes marketing ou publicitaires (les communautés de fans de produits Apple en ligne). Mais ces certitudes n'éclairent pas la question de la *forme* de la diffusion ou de la propagation dont on peut supposer qu'elle est contrainte par la *structure* du réseau. Si quelque chose devient *populaire* sur le Web, comment cela le devient-il ? Encore une fois, la nécessité de photographier la dynamique d'un processus de propagation s'impose.

C'est là que l'accès à de grandes bases de données devient capital. À moins d'être devin, il est toujours aussi difficile (voire impossible) de prévoir la diffusion à grande échelle sur le Web d'un phénomène informationnel, et donc de mettre en place par avance un dispositif d'observation. Dans certains cas, comme un événement majeur programmé (par exemple une coupe du monde de football ou des élections présidentielles, on peut s'attendre à des phénomènes massifs issus des grands médias et de la couche haute du Web qui vont « communiquer » et/ou à l'effervescence générée dans certaines communautés déjà identifiées et impliquées par nature. C'est un scénario, mais il en est d'autres à identifier : quelle histoire et quelle forme ont pu prendre, par exemple, la mobilisation des internautes contre le projet de Constitution européenne (2005), les parabènes dans les cosmétiques (2008) ou la peur de la grippe aviaire (2009) ? Prévisibles ou non, de tels événements doivent pouvoir être traduits, interprétés ou analysés en fonction des dessins que la structure leur imprime.

L'interrogation des données déjà archivées est donc aujourd'hui incontournable pour essayer de repérer un « événement » informationnel remarquable (ou une succession d'événements) et examiner son empreinte topodynamique. Et c'est le travail auquel se livre Duncan Watts dans son analyse des *motives* générés par toutes ces

marques produites par les internautes dans leurs activités quotidiennes (*tweets* et *retweets*, *like*, liens hypertextes de citations, *tags*, signalement de messages sur les plates-formes sociales ou de jeux...). Dans un premier temps, il examine différents silos d'information (dont Twitter) pour constater... qu'il ne se passe presque rien de notable (« much of the time, nothing happens »). À l'évidence, le Web n'est pas cette grande machinerie d'interactions sociales où chacun serait à l'écoute de ses voisins, et tout le temps ! La dynamique sociale attendue laisse place à ce qui apparaît plutôt comme une multitude de solitudes. Il note que dans 99 % des cas, un message émis n'est pas repris (cité, tagué, lié...) ou alors qu'une seule fois. Autrement dit, le temps ou la distance moyenne de l'empreinte topodynamique type est soit de 0 soit de 1, parfois 2. Sous cet angle, le temps sur le Web se donne d'abord comme d'innombrables microévénements qui n'ont guère de retentissement. Duncan Watts établit la synthèse des types de chemins parcourus par une information ou par l'établissement d'une connexion dans laquelle il fait figurer tous ces *little motives* qui semblent peupler majoritairement le Web social :

Image 8.3

Les « petits motifs » de connexion qui semblent peupler majoritairement le Web social, selon Duncan Watts.

Source : S. Goel, D. Watts et D. Goldstein, « The structure of online diffusion networks », dans *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*, New York, Association for Computing Machinery, 2012, pp. 623-638, <https://doi.org/10.1145/2229012.2229058>.

Tous ces motifs communicationnels extrêmement réduits, et largement majoritaires sur le Web, ne permettent pas d'éclairer de façon plus large le réseau en cartographiant des phénomènes dynamiques plus vastes. C'est pourquoi, dans un second

temps, Duncan Watts construit une étude plus étendue avec des données massives : 1,4 milliard de *tweets* associés à 65 millions de comptes et concentrés sur des milliers d'événements informationnels comme la circulation d'une vidéo, des propos politiques repris et « faisant le buzz », la relation d'une catastrophe ou d'actes terroristes de grande échelle, des rencontres sportives... L'idée que poursuit Duncan Watts est de constituer une bibliothèque de cas à partir des différents types d'empreintes topodynamiques qu'il isole dans les *data sets* (« diffusion paths in a structural way »). La synthèse qu'il en fait retranscrit les différents modes de propagation de l'information à travers les réseaux sociaux :

Image 8.4

Empreintes topodynamiques isolées par Duncan Watts.

Source : S. Goel, A. Anderson, J. Hofman et D. Watts, "The structural virality of online diffusion", *Management Science*, 62 (1), 2015, pp. 180-196, <https://pubsonline.informs.org/doi/10.1287/mnsc.2015.2158>.

Dans ce tableau, le premier et le dernier cas représentent deux extrêmes ou, si l'on veut, deux scénarios types de propagation de l'information. Dans le premier, la totalité des nœuds « touchés » (reprises, signalements, *retweets*, *tags* comme les *like* sur Facebook...) le sont directement à partir d'une source qui semble rayonner en touchant d'un seul coup son public, de façon large et massive. C'est le modèle *broadcast*, attesté ici en termes d'empreintes topodynamiques. À l'opposé, dans le dernier cas, les données valident un modèle de type *viral* avec une « contamination » de proche en proche, d'individu à individu (ou de communauté à communauté) par proximité structurale. Entre l'un et l'autre extrêmes se déclinent tous les autres scénarios qui semblent hybrider les modèles *broadcast* et *viral*. La façon dont un phénomène

informationnel se propage éclaircissant donc la structure du réseau sous différents angles, mais dans un *champ de possibles* fixé par deux extrêmes (propagation massive et globale ; propagation de proche en proche au niveau local). Les données de Duncan Watts ne disent rien sur la durée réelle des phénomènes observés, ni même d'ailleurs sur les détails de la propagation (par exemple, dans le cas du scénario *broadcast*, les nœuds « touchés » le sont-ils simultanément ou bien successivement ?). En revanche, elles indiquent clairement, dans la typologie des chemins parcourus, que les nœuds impliqués jouent différents types de rôle dans la structure globale du système. Dans le cas d'un scénario de type *broadcast*, les nœuds qui diffusent en premier l'information (les déclencheurs si l'on veut) sont topologiquement reliés à beaucoup d'autres de façon directe, correspondant à des *hubs* très connectés de la couche « haute ». Dans le cas d'un scénario de type viral, le modèle de diffusion ferait plutôt penser à la façon dont peut circuler une information dans la couche médiane des agrégats ou des communautés épistémiques qui élaborent de façon plus collective leurs savoirs, leurs valeurs ou leurs informations.

Le rapprochement que j'opère ainsi entre les données de l'observation et *le modèle en couches du Web*¹⁰ peut paraître trop rapide, presque arbitraire. Cependant j'en reste convaincu : ce modèle peut éclairer la forme du Web, mais aussi les dynamiques temporelles qui l'animent, à partir des *mêmes patterns* topologiques.

Image 8.5

Les scénarios de propagation de l'information (à droite) se rapprochent des couches du Web (à gauche).

10. Voir chapitre 4 du présent ouvrage, « Agrégats ».

Entre le modèle *broadcast* (un grand média diffuse une information reprise par une multitude d'internautes, immédiatement) et le modèle *viral* (par contamination successive des nœuds au niveau médian), tous les scénarios d'hybridation sont possibles. Et ils correspondent aux premières données analysées par Duncan Watts. *Broadcast* ou *viralité* ? La réponse est probablement entre les deux dans différentes formes d'hybridation. Par exemple, c'est le cas quand un internaute diffuse un document, repris ou cité dans sa communauté (techniquement, au sein de son *agrégat*), puis repris et cité par un ou plusieurs médias et enfin repris et cité dans de multiples agrégats sur le Web. Les cas me semblent nombreux de ce cheminement : un militant diffuse une vidéo sur les pratiques inacceptables des abattoirs en matière de respect de la vie animale, le succès local de la vidéo entraînant sa « publicisation » dans un *hub* de la couche haute (par exemple, tweeté par une agence de presse ou repéré par un journaliste pour sa réception massive sur YouTube) et, de là, commentée, reprise, citée, pointée par une multitude d'internautes dans les heures qui suivent. Et le processus inverse est tout aussi vérifiable selon moi : une campagne orchestrée par les médias et différents pouvoirs en place peut très bien ne pas trouver d'écho dans les communautés épistémiques constitutives des agrégats de la couche médiane du Web. Elle peut même engendrer des formes de rejet massif, comme en 2005 lors de la campagne pour le projet de constitution européenne soumis au référendum en France. La mobilisation des acteurs de cinéma, des sportifs, des hommes politiques, des dirigeants d'entreprise, des artistes, sur le Web et ailleurs, n'avait pas suffi à convaincre la nation. Elle avait même abouti à un contre-mouvement, qui lui-même avait généré à l'époque ses propres *hubs* (dont le site web d'Étienne Chouard, une référence depuis lors¹¹). Évidemment, le référentiel commun du temps ne permet guère de rendre compte de la nature de ce qui se joue dans un réseau comme le Web : la multitude de scénarios hybrides possibles entre *broadcast* et *viralité* se déploie-t-elle à l'échelle des secondes, des heures, des jours, des semaines ? D'une certaine façon, peu importe. Seule compte, pour le moment, la façon dont s'enchaînent les événements dont la fameuse *cascade* est l'une des figures essentielles et les plus étudiées (par exemple, les pannes en cascade dans des systèmes techniques complexes comme les réseaux de distribution de l'énergie, les télécommunications ou les grands systèmes d'information). Je suis aujourd'hui convaincu que les événements s'enchaînent dans la topologie du Web et que les données de l'observation en portent l'empreinte. Sur ce point, mon référentiel est le modèle en couches du Web que nous avons élaboré, et il semble éclairer singulièrement les figures centrales que découvrent petit à petit les chercheurs comme Duncan Watts¹².

11. <http://chouard.org/blog/>.

12. D. Watts, *Six Degrees: The Science of a Connected Age*, New York, Norton, 2004.

Image 8.6

(a) Le scénario *broadcast* avec une diffusion massive depuis la couche touchant rapidement la grande majorité des nœuds du système ; (b) Le scénario viral avec la logique de diffusion de proche en proche et d'agrégats en agrégats ; (c) Hybridation des deux scénarios types en de multiples formules.

Il paraît logique d'imputer les phénomènes *broadcast*, ou de diffusion massive sur le Web, aux différents rôles que joue la *couche haute* du réseau, la seule à générer la connectivité nécessaire à la fulgurance et à l'universalité d'une diffusion générale. On peut inscrire dans la couche haute l'ensemble des grands *hubs* du Web, soit parce qu'ils diffusent de l'information au sens des médias « verticaux » classiques soit parce qu'ils organisent la distribution de la connectivité (les fameuses plateformes du Web 2.0). Il me semble d'ailleurs que la limite entre les deux familles tend de plus en plus à se confondre, notamment parce que les plateformes avancent ou publisent certains contenus et, inversement, parce que les grands médias d'information cherchent, eux, à fédérer leurs auditoires via de nombreux instruments d'échanges (notamment, je lis souvent attentivement les commentaires des internautes suite à la publication d'articles de presse en ligne). Pour donner un aperçu de la vitalité de cette couche haute, voici les statistiques du trafic que génèrent quelques-uns des principaux « super *hubs* » de la couche haute du système :

Chaque seconde sur le Web en 2013

Google, 2 millions de recherches	Pinterest, 11 000 utilisateurs
Spotify, 14 nouvelles chansons ajoutées	LinkedIn, 11 000 recherches professionnelles
Domain Name, 70 nouveaux enregistrements	Amazon, 83 000 dollars de transaction commerciale
Apple, 15 000 téléchargements sur iTunes	Snapchat, 104 000 photos partagées
Wordpress, 347 nouveaux posts sur les blogs	Twitter, 278 000 nouveaux titres
Web, 571 nouveaux sites	Flickr, 20 millions de vues
Walmart, 17 000 transactions	Skype, 1,4 million de connexions téléphoniques
Tumblr, 20 000 nouvelles photos	Facebook, 41 000 nouveaux posts
Emails, 204 millions d'emails envoyés	YouTube, 72 heures de vidéos mises en ligne
Instagram, 3 600 photos publiées	

Source inconnue

De façon complémentaire, l'idée, voire l'idéologie, de la viralité (empruntée à l'épidémiologie dans ses modèles) me semble rendre compte de cette dynamique liée aux formes sociales d'organisation de communautés, souvent synonyme « d'engagement » de leurs acteurs. À ce niveau du Web, la circulation de l'information est alimentée par des phénomènes de proximité, des liens tissés de longue date, des proximités locales ou des réseaux de confiance en termes d'identification des sources.

On peut donc supposer que les scénarios de diffusion ou de propagation (d'une information, mais aussi d'un flux ou d'une valeur) se situent *quelque part* entre *broadcast* et viralité, et en fonction de la topologie du réseau qui imprime alors au temps ses formes propres. Le principe de l'hybridation, ce *quelque part*, relève de la nécessité structurale si l'on se réfère au modèle du Web en couches : peut-on imaginer une dynamique informationnelle se développer sans inclure à la fois l'infrastructure de grands systèmes (notamment de connexion sociale) et les propriétés locales et particulières des communautés d'acteurs regroupés en agrégats ? Là aussi, c'est l'accumulation des cas d'observation, éventuellement organisés en bibliothèque, qui permettra de développer un appareil conceptuel et méthodologique efficient pour classer les types topodynamiques « d'événements » propagatoires, les repérer dans les *data* et, peut-être un jour, apercevoir des « cycles » ou des formes de périodicité typiques jusqu'à leurs prémisses sous forme de « signaux faibles ». Mais nous en sommes loin et la difficulté de la tâche nous avertit du danger des simplifications que suppose le « marketing viral », les méthodes et les

instruments de e-réputation, du flou qui entourent certains concepts comme ceux « d'influenceurs » ou de « communautés actives ».

Les modèles théoriques balancent eux aussi entre deux scénarios types, dont l'un et l'autre sont proches du principe de la diffusion virale et de la diffusion broadcastée¹³. Le premier consiste à supposer que les idées, les innovations, les *news*, les idéologies se propagent sur le Web comme le font les virus dans une population, par contamination de proche en proche. C'est notre façon spontanée d'envisager les phénomènes informationnels, presque indépendamment de la forme du réseau qui nous relie, en empruntant aux grandes épidémies toute une imagerie impressionnante (comme celle de l'Europe contaminée par la peste noire au Moyen Âge, par vagues successives). Ce principe de contamination par contiguïté physique dans le cas des virus (biologiques) nourrit l'idée largement répandue qu'il en va de même avec l'information à travers les réseaux où compteraient, plus que tout, le nombre d'individus exposés et la surface occupée par le phénomène, dessinant les frontières d'un *disease front*. C'est là que le modèle viral, en marketing par exemple, puise quelques-uns de ses arguments, comme parfois aussi la communication politique, comme si les « agents » ou les « acteurs » ne produisaient pas d'inférences sur la forme et les méandres du réseau qui les relie. Le principe de la diffusion de proche en proche n'est pas niable, notamment à l'échelle des agrégats entre acteurs d'une communauté (ce que l'on appelle aussi les *peer-to-peer models*), mais il faut en situer la mécanique dans le cadre plus large des connexions possibles avec les communautés voisines et, surtout, avec la couche « haute » du Web.

Image 8.7

Modèle épidémique de contagion par la connectivité.

13. Parmi les nombreux articles scientifiques, on pourra se référer à : J. Coleman, E. Katz et H. Menzel, "The Diffusion of an Innovation Among Physicians", *Sociometry*, 20 (4), 1957, pp. 253-270 ; M.S. Granovetter, "Threshold Models of Collective Behavior", *American Journal of Sociology*, 83 (6), 1978, pp. 489-515 ; D. Kempe, J. Kleinberg et E. Tardos, "Maximizing the Spread of Influence through a Social Network", 2003, <https://www.cs.cornell.edu/home/kleinber/kdd03-inf.pdf>.

Au principe de la contagion (sociale, informationnelle) et de la viralité supposée de certains éléments locaux (innovations, idéologies, révélations...) qui se propagent petit à petit à l'échelle d'un système, il faut pouvoir associer des modèles d'analyse plus globaux qui rendraient compte de phénomènes massifs comme des mutations globales ou la logique d'enchaînements inéluctables comme les « cascades » (particulièrement en cas de défaillances d'un système distribué). Les jeux de données massives que l'on pourrait traiter sous forme d'empreintes topodynamiques ne sont guère accessibles aujourd'hui (du moins, à ma connaissance), leur traitement supposant par ailleurs une infrastructure de calcul significative et hors de portée du cartographe que je suis. À cette échelle, ce sont plutôt les modèles théoriques qui dominent, avec leurs cortèges de concepts issus des sciences physiques notamment comme ceux de *transition de phase*, de *seuil critique* (ou de *tipping point*) ou de *percolation*. Ces concepts, bien au-delà des sciences du Web, ont marqué l'histoire des modèles d'évolution des systèmes complexes et distribués, en particulier dans les domaines de la sécurité industrielle (le *tipping point* dans le cas des catastrophes industrielles), de la biologie (mutations génétiques rapides), de l'histoire (les révolutions), de la chimie (passages progressifs ou fulgurants des états solides-liquides-gazeux), de l'économie (crises financières). Dans une perspective de graphes appliqués à ce type de système, l'investigation scientifique repose sur l'idée de prendre en compte l'ensemble des paramètres d'un réseau ouvert et distribué (comme le Web), notamment celui du diamètre et des chemins les plus courts, celui du seuil de tolérance ou de résistance des nœuds du système (qui « filtrent » l'information et la rendent acceptable) et, enfin, l'hypothèse que l'évolution globale de vastes systèmes complexes est régie par une série d'« événements », voire de crises qui aboutissent à sa mutation, à son évolution, à son adaptation. Duncan Watts, Albert-László Barabási ou Steven Strogatz en ont étudié certains aspects sous forme de modèles abstraits et complexes mathématiquement à calculer. Duncan Watts, en particulier, essaie de concevoir des scénarios globaux de diffusion en croisant les propriétés topologiques d'un réseau (distribution des liens en loi de puissance entre les éléments, diamètres courts...) avec différents seuils de résistance pour chacun des nœuds. Dans ce type de modèle, les nœuds du graphe (donc les acteurs-diffuseurs-relayeurs d'informations) sont considérés comme des agents connaissant effectuant des inférences sur leur environnement et qui tiennent compte des externalités (comme l'« état » de leurs voisins). En effet, dans le cas de la diffusion d'informations politiques, idéologiques, religieuses ou qui touchent plus généralement à la décision et aux valeurs, le nombre de voisins d'un acteur avec lesquels il est relié comme leurs différents seuils de résistance (qui peuvent être très variables) entrent comme variables d'un modèle de diffusion (ce qu'ont noté les économistes quand ils étudient la diffusion des innovations).

Image 8.8

Plusieurs modèles de diffusion de l'information sur les réseaux s'appuient sur la corrélation des propriétés topologiques (distribution des liens, diamètre, densité, composants...) avec des seuils de résistance plus ou moins élevés (ici, ramenés à une variable 0-1). En fonction du premier nœud activé – ou point de départ dans le graphe – et de certaines propriétés de l'information diffusée à ses trois voisins directs, l'un d'eux change d'état (à droite), car il a un seuil de résistance relativement faible. Il peut donc à son tour diffuser l'information dans le réseau, diffusion qui sera elle-même dépendante du nombre de voisins immédiats et de leurs seuils respectifs de résistance. En faisant varier le départ d'activation dans un graphe ainsi que les seuils de résistance, il arrive dans certains cas que l'ensemble du système mute et se modifie.

C'est de ce type de dispositif expérimental que Duncan Watts tire par exemple son modèle de *propagation en cascade*¹⁴. Bien que dans cette démarche mathématique les réseaux soient souvent produits de façon aléatoire et non issus de données réelles, certaines conclusions qu'en tire Watts semblent précieuses. Une diffusion globale « en cascade » n'est envisageable qu'à deux conditions. La première est que l'on trouve dans le réseau un nombre suffisant de nœuds ayant des seuils de résistance peu élevés (ou, si l'on veut, qu'il y ait un public « à convaincre »). La seconde est que le nombre de liens par nœud (la densité) soit suffisamment important (pour « faire réseau », au sens par exemple des graphes de données web), mais pas trop élevé non plus. En effet, dans un univers massivement connecté (donc très dense), sachant que chacun des nœuds produit des inférences sur l'état de ses innombrables voisins immédiats, l'activation d'un seul ne suffirait pas à enclencher la moindre propagation. Par contre, quand le phénomène apparaît, il est fulgurant et massif puisque ce sont les nœuds les plus densément connectés au réseau qui entament le processus. C'est ce que les physiciens appellent une transition de phase *discontinue*. Et la description

14. Voir notamment le chapitre 8 de D. Watts, *Six Degrees*, op. cit.

que fait Duncan Watts de cette configuration abstraite semble correspondre assez bien à ce que nous avons défini comme une « couche haute » du Web. Et l'auteur poursuit en notant que plus on diminue statistiquement le nombre de liens par nœuds (diminution de la densité), plus la propagation à l'échelle du réseau prend du temps, de *cluster* à *cluster*, comme s'il s'agissait d'un phénomène de diffusion lente et latérale d'agrégat en agrégat. Au niveau le plus bas, en imaginant que chaque nœud n'est relié qu'à un seul autre, le succès ou l'échec de la propagation dépend donc essentiellement de la distribution des seuils de résistance à l'échelle du réseau.

Notre capacité à saisir la dynamique temporelle sur les réseaux dépendra donc de la façon dont les données valideront, selon les cas, des modèles de propagation locale ou globale, instantanée ou successive et toujours en articulant la forme des empreintes recueillies aux propriétés topologiques. Le principe du modèle du Web en couches indique, à lui seul, combien les figures du temps varient suivant le type de topologie hypertexte dans lequel on se trouve. C'est une piste parmi d'autres et d'autres concepts mériteraient d'être interrogés à la lumière des premières hypothèses produites par les scientifiques. Parmi eux, Steven Strogatz se distingue pour avoir fait de la synchronie une « science », interrogeant inlassablement différents systèmes distribués pour leurs propriétés évolutives¹⁵. Le point de départ de sa quête réside dans l'hypothèse qu'il n'existe pas d'horloge centrale dans les grands systèmes distribués comme les écosystèmes, le Web ou les cellules et que le temps est inscrit dans la forme du réseau étudié. En suivant Steven Strogatz dans son investigation des réseaux (dont ses travaux communs avec Duncan Watts) et de leurs différents modèles scientifiques, on comprend la fascination du chercheur pour la *synchronie*, un phénomène où toutes les dimensions d'une topologie (nœuds, liens, couches...) fonctionnent à l'unisson sans laisser apercevoir la moindre dilatation, la moindre durée. Mais il y a loin des réseaux techniques de capteurs ou de transformateurs électriques (comme souvent en recherche technologique) aux réseaux humains et à leurs interactions, numériques ou non : peut-on imaginer tous les hommes agir de façon synchrone, penser le monde d'une façon unique et se connecter de façon prévisible dans un espace social où chacun serait, en quelque sorte, à sa place ?

15. S. Strogatz, *Sync: The Emerging Science of Spontaneous Order*, Londres, Penguin, 2004.

Chapitre 9. L'information, en n dimensions

« We're still waiting for a major breakthrough in understanding [non-linear systems], and it could be a long time in coming. I think we may be missing the conceptual equivalent of calculus, a way of seeing the consequences of the myriad interactions that define a complex system. It could even be that this ultracalculus, if it were handed to us, would be forever beyond human comprehension. We just don't know », Steven Strogatz, *Sync: The Emerging Science of Spontaneous Order*, 2003.

Comme beaucoup, je suis resté rivé à mon poste de télévision en ce mois de janvier 2015. Je n'ai pas pu m'en détacher lors de l'attentat contre *Charlie Hebdo*, puis lors de la prise d'otages de l'Hyper Cacher, sidéré par les proportions que prenaient ces événements dans les médias. À vrai dire, j'étais plutôt *englouti* par le flot d'informations qui écrivait l'histoire devant moi, en direct. Ce n'est que petit à petit que je comprenais les implications de ces événements à plus long terme, qu'elles soient culturelles, politiques, légales ou... techniques. Passée la période d'émotion se profilaient les questions éthiques et politiques de la surveillance des réseaux d'activistes islamistes, de l'accès pour les services de renseignement à des données inédites, à des sources multiples et de leurs croisements, sans oublier les aspects juridiques qui allaient émerger. Pour le moment, j'étais pris dans un flot d'images et de commentaires d'où émergea subitement l'envie de noter sur un plan les faits et les personnages, un instrument de compréhension globale où se relieraient de façon logique (ou causale) les événements et que l'on pourrait inscrire à mesure qu'ils apparaissaient. Vous savez, ce genre de grand tableau synthétique qui équipe probablement les rédactions de nos médias ou les *war rooms* et sur lequel s'accumulent les informations en temps réel. Certains médias, comme France 24 (ci-dessous), en diffusèrent d'ailleurs quelques-uns où se trouvent épinglés les frères Kouachi et Amedy Coulibaly, les acteurs principaux autour desquels s'agrégent d'autres éléments qui contribuent, justement, à *faire tableau*. Mon envie se transforma rapidement en projet. Après tout, j'avais de quoi bricoler mon propre schéma, car j'avais tout sous la main et autour de moi : l'équipement qui m'accompagnait dans l'exploration du Web. Et ce petit détour du côté des graphes de réseaux terroristes ne fut pas sans conséquence théorique et méthodologique pour la quête qui m'animait depuis déjà quelques années autour de la forme du Web...

Image 9.1

Visualisation de l'enquête sur le nébuleuse Coulibaly-Kouachi.

Crédits : studio graphique France Médias Monde (<https://www.france24.com/fr/20150115-infographie-nebuleuse-coulibaly-kouachi-attentats-paris-charlie-hebdo-terrorisme-police-france>).

Je n'étais pas spécialiste du renseignement ou des questions de sécurité, mais je savais, cependant, que pour rôder une méthodologie, affiner des instruments, produire des hypothèses ou même un modèle général (ici d'un type particulier, celui d'une d'organisation sociale) il faut d'abord disposer d'une bibliothèque de cas réels. C'est alors que je suis revenu sur le travail de Valdis Krebs¹, l'un des premiers à avoir démontré tout l'intérêt du Social Network Analysis pour les questions de sécurité et de description des organisations terroristes.

Dans les jours qui ont immédiatement suivi les attentats du 11 septembre 2001, Valdis Krebs a commencé à reconstituer sous forme de graphe le puzzle des cellules qui ont

1. <http://web.archive.org/web/20210415084045/http://www.orgnet.com/hijackers.html> ; [http://insna.org/PDF/Connections/v24/2001_I-3-7 .pdf](http://insna.org/PDF/Connections/v24/2001_I-3-7.pdf).

opéré au nom d'Al-Qaïda aux États-Unis. Comme à chaque fois, tout débute avec une table de données (type fichier Excel) dans laquelle il va accumuler, rivé aux chaînes de télévision et au Web, différentes informations en agrégeant les sources comme le ferait un détective. Dans ce travail remarquable de précision et de patience, sa compréhension du phénomène se construit par élargissement progressif de son horizon d'observabilité, au fur et à mesure des versions successives de sa « carte ». Chaque nouvelle information vient nourrir la perspective : Valdis Krebs ajoute, en particulier, de *nouveaux liens* entre les 19 membres des quatre cellules qui concernent les points communs de leurs différentes biographies (certains sont de la même famille, ont fréquenté les mêmes lieux, vécu dans les mêmes camps d'entraînement en Afghanistan, sont des amis d'enfance ou ont côtoyé les mêmes imams). Il ajoute aussi différentes informations juridiques, commerciales ou administratives qui lui permettent alors de comprendre le lent travail préparatoire des attentats (location commune de voiture, logements communs, voyage à deux ou trois membres... sur une période de plusieurs mois ou années). Une partie significative de son travail a consisté à intégrer aussi de nouveaux nœuds dans la carte, tous ces acteurs qui ont assuré la logistique des opérations et/ou qui ont croisé les « kamikazes » au cours de leur vie. Le principe cumulatif adopté par Valdis Krebs le conduit donc à intégrer de nouveaux nœuds dans le système en fonction du *type de liens* auquel il s'attache.

Il s'agit là d'un principe d'expansion contrôlée du réseau sur des critères très factuels et qualitatifs, un travail d'orfèvre.

La carte finale que livre Valdis Krebs comporte ainsi plus de 60 membres (dont certains sont seulement soupçonnés d'avoir participé de façon indirecte aux quatre attentats, d'autres ont été condamnés, puis relâchés aujourd'hui, et d'autres encore à l'identité invérifiable). La plupart des quarante nouveaux nœuds de la carte ont été ajoutés à partir d'un double critère : la nature de leurs liens avec les 18 membres des 4 cellules opérationnelles et le fait qu'ils ont été en contact (pour la plupart) avec au moins deux terroristes (presque sur le même principe que la co-citation en bibliométrie).

Image 9.2

Extraits de l'enquête de Valdis Krebs sur les réseaux terroristes.

Source : V. E. Krebs, "Mapping Networks of Terrorist Cells", *Connections*, 24 (3), 2002, pp. 43-52.

Je me suis donc mis à construire mon propre jeu de données sur les événements du 11 septembre pour rejouer l'expérimentation, comme on le ferait en sciences expérimentales pour valider une « manip » et ses résultats. Les informations ont été ajoutées petit à petit dans la table de données disponible sous Gephi (une sorte de feuille Excel directement visualisable sous forme de graphe).

Image 9.3

Nodes	Id	Label	TYPES	ISO	Eccentricity	Closeness Cent.	Betweenness...	LAST
Mohamed Atta	1	Mohamed Atta	1	http://fr.wikipedia.org/wiki/Mohammed_Atta	2	1,806	562,202	Egypte
Marwan al-Shehhi	2	Marwan al-Shehhi	2	http://fr.wikipedia.org/wiki/Marwan_al_Shehhi	2	1,75	88,900	Sau
Hani Hanjour	3	Hani Hanjour	4	http://fr.wikipedia.org/wiki/Hani_Hanjour	2	1,289	126,189	Arabie Saoudite
Basel Sam Ben Ghemais	4	Basel Sam Ben Ghemais	0	http://en.wikipedia.org/wiki/Basel_Sam_Ben_Ghemais	4	2,778	0	Turisie
Nawaf Alhazmi	5	Nawaf Alhazmi	4	http://fr.wikipedia.org/wiki/Nawaf_al_Hazmi	4	2,111	69,540	Arabie Saoudite
Ramzi Bin al-Shih	6	Ramzi Bin al-Shih	0	http://fr.wikipedia.org/wiki/Ramzi_Bin_al-Shih	4	2,417	40,417	Liban
Ziad Jarrah	7	Ziad Jarrah	3	http://fr.wikipedia.org/wiki/Ziad_Jarrah	2	1,806	105,622	Liban
Abdul Aziz Al-Oman	8	Abdul Aziz Al-Oman	1	http://en.wikipedia.org/wiki/Abdul_Aziz_al-Omani	3	1,722	87,3	Arabie Saoudite
Fayez Ahmed	9	Fayez Ahmed	2	http://fr.wikipedia.org/wiki/Fayez_Ahmed	4	2,056	17,640	Sau
Salem Alhazmi	11	Salem Alhazmi	4	http://fr.wikipedia.org/wiki/Salem_al_Hazmi	4	2,388	8,386	Arabie Saoudite
Satam Suqami	12	Satam Suqami	1	http://fr.wikipedia.org/wiki/Satam_al_Suqami	4	2,167	24,933	Arabie Saoudite
Zacarias Moustouli	13	Zacarias Moustouli	0	http://fr.wikipedia.org/wiki/Zacarias_Moustouli	5	3,389	0	France
Hamza Alghamdi	14	Hamza Alghamdi	2	http://fr.wikipedia.org/wiki/Hamza_al_Ghamdi	4	2,083	17,472	Arabie Saoudite
Said Bahaji	15	Said Bahaji	0	http://en.wikipedia.org/wiki/Said_Bahaji	4	2,056	54,854	Allemagne
Khalid al-Midhar	16	Khalid al-Midhar	4	http://fr.wikipedia.org/wiki/Khalid_al-Midhar	4	2,639	0	Arabie Saoudite
Saeed Alghamdi	17	Saeed Alghamdi	3	http://fr.wikipedia.org/wiki/Saeed_al-Ghamdi	3	2,167	29,662	Arabie Saoudite
Wail Alshahr	19	Wail Alshahr	1	http://fr.wikipedia.org/wiki/Wail_al_Shehri	4	2,25	0	Arabie Saoudite
Waleed Alshahr	20	Waleed Alshahr	1	http://fr.wikipedia.org/wiki/Waleed_al_Shehri	4	2,222	1,672	Arabie Saoudite
Agus Budiman	22	Agus Budiman	0	http://912review.org/Sept12008/Budiman_Agus.shtml	4	2,194	7,028	Indonesie
Ahmed Alghamdi	23	Ahmed Alghamdi	2	http://fr.wikipedia.org/wiki/Ahmed_al_Ghamdi	4	2,028	24,244	Arabie Saoudite
Zakariya Essabar	25	Zakariya Essabar	0	http://en.wikipedia.org/wiki/Zakariya_Essabar	4	2,528	0,783	Maroc
Ahmed Al Hazzani	26	Ahmed Al Hazzani	3	http://en.wikipedia.org/wiki/Ahmed_al_Haznawi	3	2,056	14,51	Arabie Saoudite
Imad Eddin Bari	27	Imad Eddin Bari	0	http://en.wikipedia.org/wiki/Imad_Eddin_Bari	4	2,528	3,672	Espagne

Table de données Gephi.

J'ai ainsi produit un graphe de 37 acteurs dont 19 ont constitué les 4 groupes kamikazes. J'ai repris une partie des informations à l'étude de Valdis Krebs, mais j'en ai rajouté d'autres empruntées aux biographies, parfois très détaillées, des terroristes sur Wikipedia². Quelques-uns parmi eux ont une identité supposée, d'autres sont toujours en prison aujourd'hui. C'est pour cela que je n'ai retenu que 18 acteurs associés aux équipes opérationnelles des attentats, selon différents aspects comme le rôle joué par certains dans le soutien logistique ou le fait que d'autres sont des compagnons sur les différents terrains d'action du terrorisme islamiste (Afghanistan, Pakistan, Irak). Il est important de noter que, quand on étend le graphe des relations au-delà de ces 18 acteurs associés, on arrive rapidement à des réseaux issus d'Al-Qaïda ou de l'État islamique dans lesquels ont gravité les frères Kouachi et Amedy Koulibaly. En deux « clics » ou en deux accointances sociales de distance, on peut donc construire une vaste carte où serait présente une grande partie des réseaux islamistes actifs aujourd'hui. À une condition cependant : de faire porter sur le processus d'expansion du graphe ou de la carte des critères qualitatifs rigoureux (séjours communs dans un camp d'entraînement en Afghanistan, fréquentation de la même mosquée radicale...). On imagine aisément à quelle situation absurde mènerait une expansion générale où pourrait figurer au final une bonne partie de l'humanité ! Après tout, nous sommes tous liés les uns aux autres, directement ou indirectement, d'une façon ou d'une autre, ne serait-ce que sur Facebook ou Gmail. Le danger du *6 degrés* et des proximités structurales³.

Je dois avouer que j'entrais là dans un univers qui s'est vite révélé différent de celui de l'exploration du Web et de ses masses de données. Dans le travail que j'entamais et où une grande attention est portée à la vérification des informations pour un graphe comportant seulement 34 acteurs, on prend rapidement conscience en le construisant de l'amoncellement des difficultés méthodologiques. Le travail qualitatif de l'enquête (la mienne étant fictive) repose en grande partie sur l'analyse de la richesse et de la diversité des connexions de type social. Un autre aspect problématique de la démarche est d'assumer, pas à pas, un horizon d'observabilité du phénomène de plus en plus élargi et complexe. La prise de décision de faire entrer ou non un acteur ou un lien dans le graphe relève d'un exercice difficile entre le trop grand nombre d'informations potentiellement pertinentes et la nécessité de les qualifier précisément. L'une des grandes différences avec les univers du *data mining* et de son ingénierie parfois massive est d'assumer une forme d'incomplétude générée par la nature des

2. L'un des sites les plus utiles aura été sans conteste le site du programme START (National Consortium for the Study of Terrorism and Responses to Terrorism), à ma connaissance une référence incomparable. Parmi les projets du programme START, figure par exemple un projet de base de données de profils du radicalisme issues de l'analyse de milliers de sources d'information (projet PIRUS - Profiles in Individual Radicalization in the United States).

3. Voir chapitre 2 du présent ouvrage, « Questions de distances ».

relations causales des univers sociaux qui sont constitués d'événements, de faits ou d'aléas historiques. Contrairement au lien hypertexte qui indique de façon univoque et universelle l'existence d'une connexion ou d'une relation, mais ne contient pas, en quelque sorte, son explication. Il faut de surcroît ajouter à l'exercice cette difficulté qui consiste à opérer dans un univers temporaire et aux frontières floues : certains éléments (nœuds ou liens) intégrés peuvent se révéler par la suite inutiles et/ou injustifiés ou d'autres encore peuvent être utiles à un moment et inutiles à un autre. Quand il s'agit d'acteurs ou de groupes sociaux, on comprend les implications éthiques ou morales de la cartographie ou, dans des techniques plus quantitatives, la production de *false positive* (individus innocents répondant cependant aux critères d'identification des algorithmes en *big data*). Pour couronner le tout, il apparaissait que les éléments cartographiés étaient mis en relation entre eux par des liens de différentes durées (longue, courte, occasionnelle) et de différentes intensités (connexion occasionnelle dans un *meeting* politique ou religieux, connexion logistique quand deux acteurs habitent au même endroit près des lieux des attentats, connexion forte et ancienne quand on appartient à la même famille ou quand on a passé des mois dans un camp d'entraînement en Afghanistan).

Pour toutes ces raisons, une carte finale ou définitive en matière de sécurité n'est guère envisageable, sauf à la construire de façon rétrospective et/ou en ayant connaissance de toutes les dimensions d'une activité criminelle. L'intérêt des cartes de réseaux criminels réside plutôt dans la dynamique de leur construction, à mesure qu'adviennent les éléments de connaissance qui permettent de comprendre un contexte. Elles sont proches des instruments de *mind mapping* ou de cartographie sémantique et sont très contrôlées qualitativement. Ce sont de véritables instruments de *dynamic knowledge discovery*, mais d'une nature différente de celles qui accompagnaient l'exploration des structures du Web et de ses *patterns* statistiques massifs. La construction de ma carte générale a débuté par l'intégration des 19 kamikazes du 11 septembre 2001, au moment où les membres des 4 cellules ont été identifiés (Valdis Krebs a commencé sa carte à partir des deux premiers suspects identifiés).

Les 4 cellules terroristes identifiées par Valdis Krebs.

À ce stade ce sont quatre *clusters* indépendants (plus haut), un pour chacun des avions (vol AA11 en rouge s'écrase sur la tour 1 du World Trade Center, vol AA77 en violet s'écrase sur le Pentagone, vol UA175 s'écrase sur la tour 2 du World Trade Center, vol UA93 s'écrase en Pennsylvanie alors qu'il se dirigeait vers le Capitole). Dans ce type de travail, les attributs associés aux liens sont d'une grande importance : ils permettent de faire apparaître des labels sur les liens, mais, surtout, de qualifier et de pondérer de nombreuses familles potentielles de connexions. La question d'un travail tant qualitatif que quantitatif sur les connexions est un aspect central de la cartographie des réseaux criminels. Pour commencer, je n'ai donc que les liens « logistiques » assurant la présence simultanée des membres du réseau dans les quatre vols différents. Ces quatre *clusters* sont isolés et ne constituent pas, à proprement parler, un réseau cartographiable : l'adjonction de liens supplémentaires permet, dans un second temps, de faire apparaître un graphe connexe où chaque acteur se trouve lié aux autres via la découverte de différents types de relations. Il aura fallu des mois aux enquêteurs pour les identifier à partir de très nombreuses sources d'information que Valdis Krebs résume sous forme d'un tableau.

Tableau des sources de données utilisées par Valdis Krebs

Relationship/Network	Data Sources
1. Trust	Prior contacts in family, neighborhood, school, military, club or organization. Public and court records. Data may only be available in suspect's native country
2. Task	Logs and records of phone calls, electronic mail, chat rooms, instant messages, Web site visits. Travel records. Human intelligence: observation of meetings and attendance at common events
3. Money & Resources	Bank account and money transfer records. Pattern and location of credit card use. Prior court records. Human intelligence: observation of visits to alternate banking resources such as Hawala
4. Strategy & Goals	Web sites. Videos and encrypted disks delivered by courier. Travel records. Human intelligence: observation of meetings and attendance at common events

C'est le principe des *layers* de liens, autrement dit des *couches* de liens en fonction de leur nature en termes de sources. Étant donné mes instruments et mes connaissances, j'ai réduit le principe à seulement deux familles (liens *forts* représentant des liens de famille, d'amitié, de scolarité commune ou de présence conjointe dans des camps d'entraînement, notamment en Afghanistan ; liens *faibles* relevant de questions qui me sont apparues plutôt logistiques et opérationnelles comme les cours

communs de pilotage, les locations de voitures et de logements, les flux bancaires). Comme le fait Valdis Krebs, j'ai attribué des forces différentes à chacune des deux familles de liens (liens faibles = poids 1, liens forts = poids 5), selon une répartition *a priori* assez arbitraire. Les formules de traduction peuvent être nombreuses, tout comme la grammaire générale qui préside à la construction des liens. Par exemple, le poids des liens peut, ou non, être cumulé comme cela est fait par Valdis Krebs (un lien faible entre A et B de poids 1 + un lien fort entre A et B de poids 5 = un lien de poids 6). Un tableau plus général se dessine alors où les 19 kamikazes sont effectivement reliés entre eux :

Les liens épais représentent les liens forts sous Gephi. Ils permettent de relier les nœuds du graphe en fonction de relations « profondes » d'amitié et d'appartenance familiale.

Quand on dispose ainsi d'un graphe connexe, on peut alors déployer les métriques de graphes qui ont guidé jusqu'à présent l'exploration des données web. Les attentats du 11 septembre 2001 peuvent *enfin* être considérés sous l'angle d'un réseau de relations sociales, historiques ou factuelles. Ce réseau possède donc un diamètre, une densité, des éléments hiérarchisables selon des scores de degrés ou regroupables en *clusters*. Spatialisée en Force Atlas 2, la structure de réseau permet par exemple d'apercevoir les rôles proéminents de Mohammed Atta, mais aussi de Marwan al-Shehhi ou de Hani Hanjour.

Image 9.7

La pondération des liens dans la table de données permet alors de faire varier les vues sur les données avec l'algorithme Force Atlas 2 (fonction « poids des liens » activable sous Gephi). De gauche à droite, trois vues sur la même structure de graphe en augmentant petit à petit la prise en compte des scores de pondération. La vue de gauche est plutôt « fonctionnelle » tandis que celle de droite a tendance à regrouper les acteurs selon leurs liens d'amitié et familiaux.

Toute la batterie de métriques réseau peut être déployée : on peut ainsi « Page-Ranker » les acteurs ou encore (comme plus haut) faire varier les forces d'attraction entre les nœuds ou les acteurs du réseau en fonction du poids des liens. Parmi toutes ces métriques, une m'interpellait : le diamètre du graphe (ramené ici à la longueur moyenne des chemins possibles entre les nœuds considérés par paire), autour de 4,5, ce qui paraît important pour un réseau de seulement 19 acteurs. Cette longueur apparente peut être considérée comme un principe de sécurité pour les membres du réseau (chaque élément étant éloigné des autres, il ne dispose, par exemple, que d'informations partielles sur l'ensemble de l'organisation). Mais cette longueur relative contredit un principe nécessaire de coordination (donc de distance moyenne *courte*) dans la préparation des attentats comme dans la synchronisation des équipes. Sous un autre aspect, des métriques comme les formes de centralité (notamment *betweenness*) indiquent le rôle prépondérant de certains acteurs, comme Mohammed Atta, qui ont pu jouer le rôle d'animateur de l'ensemble des 4 cellules. Sachant que ce dernier était aussi l'un des pilotes, ce djihadiste concentre une double compétence (*skill*) qui rend fragile l'organisation (son arrestation avant les attentats aurait peut-être permis de les éviter). L'étude de la distribution de *skills* nécessaires à l'échelle d'une organisation pour préparer et perpétrer des attentats semble être un enjeu important de programmes comme START et semble issue, pour partie, des méthodes d'analyse des réseaux sociaux. Elle nous incite à comprendre, notamment, comment la couche des *skills* permet de relier organiquement des individus à une organisation et comment elles sont distribuées depuis les équipes opérationnelles jusqu'aux cercles lointains du soutien logistique ou idéologique. Ainsi, différentes métriques permettent d'explorer le corpus sous différents angles. Par exemple, on peut comparer la répartition des acteurs tels qu'ils se sont distribués en cellules opérationnelles dans les quatre avions et la façon dont un algorithme de

détection de communautés comme Modularity les classifie. Dans la vue suivante, l'algorithme n'a pas été contraint en nombre de classes (il produit donc une distribution « optimale » en termes statistiques), mais certains acteurs forment désormais des groupes nouveaux, très liés apparemment à des questions de relations familiales, de nationalité et de coparticipation à des périodes d'entraînement en Afghanistan.

Image 9.8

À gauche, la répartition par groupe telle que la révèle un algorithme de détection de communautés comme Modularity en prenant en compte le nombre, la variété et le poids des nouveaux liens qu'ont révélés les enquêtes. À droite, la même structure, mais les couleurs et les groupes sont ceux des cellules opérationnelles. On peut ainsi considérer, à certains égards, que la mise en place de cellules opérationnelles s'appuie sur des connexions plus anciennes. En mixant les acteurs en fonction de leurs compétences (les cellules opérationnelles) et non pas de leurs accointances sociales ou de leurs connexions passées, on rend difficilement repérable une cellule active, composée a priori d'acteurs ayant des horizons différents.

Si l'on projette cette fois les 37 acteurs intégrés dans le jeu de données, une perspective plus globale s'ouvre pour la compréhension de qui fait réseau et qui imprime à une organisation sa morphologie particulière (peut-être comme une sorte d'empreinte de sa structure).

Image 9.9

Carte des 37 acteurs.

Le graphe laisse maintenant apparaître quelques propriétés nouvelles intéressantes. En termes de métrique, le diamètre se situe désormais autour de 2,5, bien que le nombre d'éléments ou d'acteurs cartographiés ait doublé. L'univers des membres d'Al-Qaïda semble plus réduit (effet *small world*), univers dans lequel peut alors s'exercer une action de synchronisation, de coordination ou d'animation de l'ensemble du réseau à partir de distances plus courtes. Si l'on poursuit à la lumière des quelques enseignements issus de l'exploration du Web, ce type de propriétés annonce aussi un effet de hiérarchisation autour d'éléments qui semblent jouer un rôle central, soit parce qu'ils développent de nombreux liens avec tous les autres nœuds du système (les *hubs*) soit parce qu'ils occupent une « place centrale » dans le graphe (différents scores de centralité). Cette dernière propriété est intéressante à mobiliser pour identifier des faiblesses potentielles de ce type d'organisation : certains acteurs jouent un rôle déterminant en termes de structure non pas parce qu'ils sont les plus connectés (voire les plus « actifs » ou « en pointe »), mais parce que c'est par eux que passe un grand nombre de connexions (chemins possibles) entre tous les éléments.

Le principe de l'extension du corpus laisse apparaître de nombreuses difficultés, et quelques « surprises ». Par exemple, l'apparition dans le graphe de Zacarias Moussaoui, de nationalité française et emprisonné à vie aux États-Unis. Son cas pose d'intéressants problèmes d'intégration au corpus : a priori, rien n'est encore venu prouver sa participation aux attentats du 11 septembre (déjà emprisonné au moment des faits), mais il est lié de bien des façons à plusieurs acteurs de la carte (Afghanistan, séjours en Grande-Bretagne...). Autre « singularité » associée au principe

de l'expansion du corpus : Valdis Krebs fait apparaître dans ses graphes la présence de Djamel Beghal, le Franco-Algérien, parfois présenté comme le « mentor » des frères Kouachi (Chérif avec qui il était en prison et qu'Amedy Coulibaly a visité en prison). Les trois se sont retrouvés par la suite dans le Cantal où Djamel Beghal a été assigné à résidence durant plusieurs mois. Je ne sais pas si les attentats récents sur notre territoire (*i.e.* janvier 2015) représentent un « 11 septembre français », mais, en termes de cartographie, nous ne sommes ainsi qu'en « distance 2 » avec les événements du 11 septembre 2001 aux États-Unis. Une cartographie du *small world* du terrorisme est donc possible, à supposer que soit contrôlé précisément le processus d'expansion du graphe.

* * * * *

Durant les mois qui ont suivi les attentats du 11 septembre, les chercheurs (dont Valdis Krebs) et les spécialistes du renseignement aux États-Unis ont tiré quelques leçons essentielles dans la lutte qui s'engageait à grande échelle contre le terrorisme, particulièrement en matière de renseignements. Et ce sont des *hard lessons*. La première de ces leçons est qu'il est possible d'agréger des informations largement publiques pour produire des instruments pertinents d'observation des organisations criminelles. Le Web fournit aujourd'hui suffisamment de données accessibles pour isoler des frontières d'organisation et comprendre la nature de leurs activités. Un jour, peut-être, verra-t-on surgir un *djihad leaks* issu d'un travail de fourmi réalisé par quelques journalistes motivés ? Une autre leçon, complémentaire, est de type juridique, voire éthique. L'opération technique d'agrégation ouvre sur la question de l'accès à de nombreuses sources d'information ou du droit à les archiver et à les croiser. La nécessité d'un accès presque total à toutes les sources d'information (auxquelles il faudrait ajouter aujourd'hui la géolocalisation des objets mobiles, Twitter, les bases biométriques...) fait figure d'évidence, ne serait-ce qu'à la lumière de notre petite expérimentation cartographique en suivant les pas de Valdis Krebs. C'est toute la logique du Patriot Act et de la mise en place aux États-Unis du Intelligence Reform and Terrorism Prevention Act en 2004. Mais elle inaugure, aussi, une frénésie de l'archivage tous azimuts dont la NSA semble être l'instrument principal. Comme beaucoup, j'ai été surpris de l'ampleur du système d'« écoutes » déployé durant des années par la NSA, telle qu'elle apparaît dans les révélations d'Edward Snowden, sans oublier, pourtant, cette impérieuse nécessité du rassemblement de l'information que produisent les individus ou les organisations sociales.

De là, une troisième leçon issue du 11 septembre 2001, plus tardive mais tout aussi importante sur le plan méthodologique : la nécessité de développer des démarches qualitatives de contrôle des masses de données. En 2006, Jeff Jonas et Jim Harper reviennent sur les méthodes déployées en matière de renseignement pour en souligner les déficiences qualitatives :

The details of the 9/11 story reveal that federal authorities had significant opportunities to unravel the 9/11 terrorist plot and potentially avert that day's tragedies. Two of the terrorists who ultimately hijacked and destroyed American Airlines flight 77 were already considered suspects by federal authorities and known to be in the United States. One of them was known to have associated with what a CIA official called a "major league killer." Finding them and connecting them to other September 11 hijackers would have been possible –indeed, quite feasible– using the legal authority and investigative systems that existed before the attacks [...] What the 9/11 story most clearly calls for is sharper focus on the part of our national security agencies and *the ability to efficiently locate, access, and aggregate information about specific suspects*. Investigators should use intelligence to identify subjects of interest and then follow specific leads to detect and preempt terrorism [...] They lacked this focus and capability –with tragic results.

Le focus qualitatif appliqué aux masses de données conduit ainsi naturellement à adopter un modèle *subject based*, en orientant « en profondeur » l'observation d'un sujet ou d'un groupe de sujets. C'est le principe d'une sorte de balance entre l'approche qualitative (celle des enquêtes de terrain, par exemple) et l'approche quantitative où le *big data* est mobilisé à la recherche de *patterns* spécifiques associées aux organisations criminelles et/ou terroristes. La traçabilité du social suppose que l'on extrait des masses de données ces *smart data* qui viennent s'agréger en tableaux cohérents et précis. Le *small* ou le *smart data* pourraient permettre d'éviter la production de ces *false positives* (personnes innocentes dont le profil a été identifié statistiquement comme suspect possible) qui alimentent encore aujourd'hui le débat aux États-Unis. Comme l'écrit Erik Dahl en 2013, « It's not Big Data, but *Little Data*, that prevents terrorist attacks », non pour revenir aux méthodes intégralement artisanales de l'enquête, mais bien pour signaler qu'en matière d'intelligence des données appliquées à la sécurité, c'est dans le croisement des deux méthodes que réside le potentiel prédictif des instruments ou des démarches (disons, de démarche *instrumentée* d'observation).

L'équilibre de la balance quali-quantitative qu'appellent de leurs vœux les spécialistes du renseignement et de la sécurité n'est qu'un aspect local et particulier

(et ô combien instructif !) de l'hybridation des deux dimensions du traitement de l'information dans une situation où tout dépend de la façon dont sont distribuées dans un système d'information des phases et des approches qualitatives et qualitatives, depuis les algorithmes jusqu'aux interfaces. C'est ce que j'appelle *l'alchimie quali-quantitative* où alternent d'un côté la production qualitative de profils d'entités qui doivent être nourris et de l'autre celle de données quantitatives pour valider certains de leurs traits. De façon complémentaire, une approche massive de type *big data* en termes de recueil et d'archivage des données doit être nourrie de *patterns* (donc de profils) élaborés qualitativement pour optimiser les recherches (comme une sorte d'ontologie sociale où seraient archivés les entités nommées, mais aussi le nombre et la nature de leurs connexions). Ces principes généraux résonnent de façon particulière dans le contexte des attentats en France, notamment sur les plateaux de télévision où nombre d'experts ont souligné tour à tour le « manque d'analyse » et l'impossibilité pour les enquêteurs d'avoir accès à de nombreuses sources d'information.

Les nouvelles lois sur la sécurité maintenant envisagées par les pouvoirs publics permettront, peut-être, d'apercevoir plus précisément les frontières et la morphologie des organisations criminelles. Ce type de recherche me paraît aujourd'hui constituer le terrain d'expérimentation le plus riche et le plus prometteur pour ses retombées sur les sciences humaines dans leurs différents questionnements du social, surtout en sociologie et en anthropologie. Mais ce terrain sécuritaire représente aussi un formidable espace d'innovation en termes d'ingénierie pour les difficultés méthodologiques qu'il suppose avoir levées. Et les enjeux sont importants puisque ces recherches sont censées mener de l'observation à la *prédiction* et à la *prévention*.

* * * * *

L'espace occupé par cette recherche paraît maintenant assez balisé et portera massivement ces prochaines années sur le renouvellement des hypothèses, des concepts, des méthodes et des instruments dédiés à l'observation et à la modélisation du social et de ses entités interagissantes. Il m'apparaît clairement aujourd'hui que ces recherches sur les *data* liées aux questions de sécurité en constituent l'avant-garde, là où va se jouer notre avenir en termes de sécurité collective. Le chemin à parcourir pour les spécialistes du domaine est encore long : leur point de départ est analogue à celui des *web scientists* au début des années 2000, coincés entre deux façons d'envisager le terrorisme qui rappellent le *nuage* et les *abîmes* du *deep web*.

D'un côté, notre petite expérience de cartographie qualitative montre que la théorie du « loup solitaire » ou du groupe isolé ne peut tenir longtemps, comme à chaque fois lorsque l'enquête se développe. Les frères Kouachi et Amedy Koulibaly sont rapidement pris et associés dans un réseau de connexions ou de relations qui contextualisent (et parfois peuvent en partie expliquer) les opérations terroristes de ce mois de janvier 2015. Elles sont nombreuses et complexes, un long travail d'identification commence. Cependant, d'un autre côté, la recherche de *patterns* sociaux, donc d'un principe d'organisation, permet de dépasser la notion de « nébuleuse djihadiste » dont parlent tant nos médias, comme si l'on avait affaire à une masse sociale indistincte. Ce qui est certain, et qui nourrit cette idée de la nébuleuse, c'est que l'on ne peut pas réduire, ni même appréhender, les réseaux djihadistes comme des organisations représentables sous forme d'arborescence (il suffirait alors dans ce cas d'en éliminer le *leader* ou les éléments des niveaux inférieurs de hiérarchie). C'est entre l'idée du *loup solitaire* et celle de la *nébuleuse* que se loge la recherche à venir, tout comme la recherche d'un modèle ou une carte du Web se place entre le nuage et le *deep web*. C'est là, dans cet espace, qu'il s'agit de placer quelques 5 000 suspects présents sur notre territoire (si l'on en croit le ministre de l'Intérieur à la fin du mois de janvier 2015), ainsi que le réseau de leurs relations directes ou indirectes. A-t-on commencé à répertorier toutes ces entités nommées et à recenser les différents types de connexion qui les unit probablement ?

En regardant la tablette Androïd dans mon salon, je me dis que si nous laissons spontanément autant d'informations personnelles aux géants de l'information et à leurs services *marketing*, nous pourrions aussi le faire pour nos enquêteurs. Parmi toutes les approches possibles, celle par graphe serait diablement intéressante pour documenter ma vie personnelle. Il pourrait en être de même pour des suspects identifiés à condition de ne pas se laisser happer par la logique foisonnante des liens qui nous unissent tous via les technologies, ou non d'ailleurs. Dès décembre 2001, Thomas Stewart soulignait dans son article "Six Degrees of Mohamed Atta"⁴ les implications sociales ou culturelles de ce que signifient les graphes, les calculs de diamètre ou la recherche de *clusters* avec des métriques de graphes. Les mesures de distance dans l'espace social, rappelons-nous, peuvent être considérées comme des indices de *proximité*, mais aussi d'*éloignement*. Pour le cartographe que je suis, le plus frappant est de prendre conscience à quel point la cartographie du social englobe potentiellement aussi bien ceux qui ont tout de l'altérité radicale comme les djihadistes que sa propre famille ou ses amis. De ce point de vue, nous sommes tous dans la *même carte*.

4. <http://faculty.cbpp.uaa.alaska.edu/afgjp/PADM610/Six%20Degrees%20of%20Mohamed%20Atta.pdf>.

Chapitre 10. La prochaine étape : la multidimensionnalité

Les enseignements que l'on peut tirer de ces expérimentations autour de la morphologie organisationnelle des cellules terroristes sont nombreux, mais aussi, à certains égards, rédhibitoires. Durant les mois qui ont suivi mes réflexions sur la cartographie des organisations terroristes (à partir de données connues), j'ai poursuivi le travail d'investigation, en faisant porter mon attention aussi bien sur les façons de traiter en amont les données que, en aval, de les représenter. Sans m'en rendre vraiment compte, je commençais à imaginer les contours d'un instrument d'observation « du social » considéré comme un ensemble de relations de nature diverse, représentable sous forme de graphe, et exhibant des propriétés réseau identifiables (dont certaines, on peut l'espérer, seront particulières aux organisations criminelles ou terroristes). Si l'on s'en tient à la cartographie de réseaux terroristes et à ses procédures, on peut légitimement penser que, une fois mise au point la méthode, le principal défi réside dans nos capacités quantitatives de récolte et de traitement de l'information. La puissance des machines et des calculs serait l'horizon à atteindre, le *big data* et le *cloud computing* en faisant rêver plus d'un. De là, l'idée que des algorithmes pourront bientôt remplacer les mécanismes de l'analyse et contribuer de façon prédictive à l'identification des entités sociales criminelles. Il devrait donc être possible de développer un système d'accès à des sources multiples d'information (souvent massives) qui pourrait suivre en temps quasi réel quelques milliers de suspects. On pourrait ainsi agréger à grande échelle de vastes silos d'informations, surtout si de nouvelles lois permettent juridiquement d'en étendre les sources. De toute évidence, l'association des *data* et des graphes dans l'observation du social (individus ou organisations, criminelles ou non) peut sembler un terrain enthousiasmant et prometteur d'expérimentation. La perspective, au moins théorique, de mettre en place un vaste système d'information depuis les *data*, mises en réseau et révélant les fameux « *hidden patterns* » m'a effleuré l'esprit pendant quelque temps, comme si l'approche cartographique allait se déployer naturellement dans le domaine de la sécurité et de l'observation du social.

Oui, mais voilà : le processus d'agrégation de l'information multisourcée (ou des informations de types multiples) ouvre sur une problématique que les technologies actuelles ne peuvent encore résoudre. Ce verrou est celui de la multidimensionnalité de l'information que nos modèles théoriques comme nos instruments techniques ne peuvent traiter qu'en surface, de façon approximative ou parcellaire. Si l'on revient

sur le travail de Valdis Krebs, ou à mon modeste essai de cartographie du réseau d'acteurs du 11 septembre, j'espère que chacun s'est aperçu de ce « tour de passe-passe » méthodologique qui a consisté à fusionner différents types de liens, en traduisant cette diversité sous forme de poids sur le principe de l'addition (liens faibles = 1, liens forts = 5). Le procédé, l'astuce plutôt, est nécessaire techniquement : les méthodes et les outils forgés pour explorer le Web ont été conçus pour extraire des *patterns* organisationnels unidimensionnels liés à la distribution des seuls liens hypertextes. Leurs limites apparaissent lorsqu'il s'agit de réseaux plus complexes à construire, surtout un réseau social d'« entités nommées » dont les dimensions sont multiples. Si A est lié à B via une série d'appels téléphoniques, A et B peuvent être *par ailleurs* reliés via un séjour en Afghanistan, la fréquentation d'une même mosquée et être amis sur Facebook. Je peux cartographier A et B en agrégeant ces quatre types de liens, en en additionnant seulement deux ou même trois, voire n'en retenir qu'un seul, mais *je ne peux en déduire des règles d'interaction entre ces différents niveaux*. Dans tous les cas de figure, mon espace logique et technique de traitement de l'information reste fondamentalement unidimensionnel. La prise en compte de la multidimensionnalité supposerait que l'on puisse inférer des enchaînements typiques de causalité : en quoi les liens de type 1 (fréquenter une même mosquée) produisent-ils des effets sur les liens de type 2 (être amis sur Facebook) ?

En apparence, et seulement en apparence, la question de la multidimensionnalité des informations sous forme de graphe paraît constituer une sorte d'amélioration naturelle de la méthode : il s'agit de considérer les liens sous l'aspect de *layers* différents (ou différents types de connexions). Visuellement, on peut superposer toutes les couches ou *layers* et les considérer indépendamment les unes des autres, ou bien en les additionnant. Pour chacun des éléments d'un système ou d'un réseau, il s'agit donc d'enrichir en précision notre capacité d'observation des connexions en introduisant le principe des *layers* dans les modèles de données pour décomposer un graphe :

Image 10.1

Couches de liens.

Dans un dispositif d'exploration de données relationnelles comme Gephi, on peut trier les liens ou les nœuds par famille et les faire apparaître tour à tour ou selon des formules d'agrégation plus complexes. Le principe des partitions peut donc être déployé aussi bien sur les liens que sur les nœuds d'un graphe. Si l'on reprend l'exemple désormais classique du jeu de données des *Misérables* disponible avec Gephi, on peut distribuer des types (L1, L2, L3) à l'ensemble des connexions de façon aléatoire et les faire apparaître comme autant de *layers* manipulables et autonomes :

Image 10.2

Séparation des couches de liens par fonction.

On peut ainsi isoler les informations d'un *layer* de graphe et calculer dans un espace autonome de nouvelles propriétés locales liées au sous-graphe produit.

Image 10.3

Hiatus méthodologique entre la présence (a) ou l'absence (b) des nœuds non connectés par la couche de liens étudiée.

C'est dans le rapport entre les vues a) et b) que réside le hiatus méthodologique qui rend d'un coup très relatif l'ensemble du travail de cartographie « *mono-layer* » tel que je le pratique depuis longtemps. Dans mon exemple des personnages des *Misérables*, les données d'ensemble sont *déjà connues* et je procède pour la démonstration

par soustraction (a), laissant apparaître l'ensemble des nœuds (ou des acteurs) du réseau. Mais imaginons la situation inverse : j'observe une organisation en réseau à partir d'un type de liens (b) ouvrant sur une réalité tronquée par rapport à l'ensemble des nœuds et des liens potentiels, que révéleraient d'autres *layers* de liens (éventuellement puisés dans d'autres sources). Que pouvons-nous saisir d'un phénomène à partir du graphe b ? Peu de choses, évidemment, à moins d'assumer que l'on peut cartographier une organisation criminelle sur la seule base des connexions via la téléphonie mobile ou la fréquentation d'une même mosquée. Mais il faut aller plus loin : qu'ai-je saisi du Web et de sa géographie en mobilisant la distribution des liens hypertextes et non pas tous ces *layers* de liens que constituent aujourd'hui les *like*, les connexions entre amis sur LinkedIn ou Facebook, les liaisons produites par la recherche de nos Copains d'avant, les annonces auxquelles j'ai répondu sur Le Bon Coin ? S'il s'agit d'éclairer un système, ou l'usage d'un système, les graphes unidimensionnels peuvent suffire quelque temps. S'il s'agit de viser au-delà les entités sociales qui l'alimentent en information, en documents, en idées, en rumeurs, en tendances, en organisations ou en groupes, alors l'entreprise n'en est qu'à ces balbutiements techniques et méthodologiques.

Pour un cartographe comme moi, la question des réseaux *multilayers* constitue ce genre de verrou scientifique que l'on peut considérer comme la prochaine étape de la cartographie d'informations. Et la cartographie des réseaux terroristes le démontre de façon exemplaire. Désormais, à bien y réfléchir, le principe des réseaux et graphes *multilayers* interroge directement la cartographie des phénomènes sociaux au sens large et pose de nombreuses questions sur les futures méthodes d'observabilité du social via le traitement de données. Par exemple, peut-on encore rendre compte d'un « objet » social (individus, organisation, jeux d'acteurs, négociation, controverses, influence...) seulement à partir des *tweets*, des liens hypertextes, des *like*, des commentaires, des recommandations, des réseaux d'amis ou des indices de réputation ? Si l'on vise à décrire des entités sociales (individus, communautés, groupes, institutions...) ou des phénomènes sociaux (diffusion, rumeurs, opinions, idéologies...), il faudra donc hybrider les sources et les dimensions pour explorer leur combinatoire à l'aide d'une forme de technologie qui n'est pas encore née. Mais les implications sont encore plus larges et emportent avec elles le projet de cartographie du Web comme système distribué. La question de la multidimensionnalité en est même devenue centrale et stratégique pour l'observation et le contrôle de la plupart des systèmes distribués à grande échelle que l'on qualifie souvent de « complexes ». C'est ici que les problématiques de cartographie

(du social, du Web, de l'information...) rejoignent un courant majeur de la recherche actuelle, l'étude des réseaux complexes et des NoN (Networks of Networks) dont on devine les implications pour la sécurisation des systèmes comme les réseaux de transport ou de distribution de l'énergie et de l'information. Un rapport européen récent¹ souligne ainsi les enjeux scientifiques et technologiques de cet immense champ de recherche qui s'ouvre aujourd'hui :

It is usually recognized that the traditional network representation is an oversimplification of the corresponding system; for instance, social interactions seldom develop on a single channel, and more than one relationship can bind pairs of people. It is thus not surprising that multilayer networks have raised a great expectation, as they should be able to better represent the topology and dynamics of real-world systems [...] The new "degrees of freedom" in multilayer systems also yield new phenomena that cannot occur in single-layer systems [...] Such a change of paradigm, that was termed in disparate ways (multiplex networks, networks of networks, interdependent networks, hypergraphs, and many others), led already to a series of very relevant and unexpected results, and we firmly believe that: i) it actually constitutes the new frontier in many areas of science, and ii) it will rapidly expand and attract more and more attention in the years to come, stimulating a new movement of interdisciplinary research.

De sérieux défis s'annoncent aussi bien pour le cartographe que pour le mathématicien dans ses calculs sur les structures de graphe, pour l'informaticien dans sa construction des modèles de données ou pour le *designer* d'interface qui doit jongler désormais avec tous les prismes d'un objet pris dans différents « ordres de rationalité » ou « plans de la réalité ». Et ces défis sont tels qu'un sentiment légitime d'impuissance peut nous habiter, doublé de prise de conscience des limites de nos outils et de nos méthodes, dont la cartographie d'information avec les modèles de données sur lesquels elle est fondée, les métriques de graphes et les figures de spatialisation qui en sont l'ossature. Face aux problèmes méthodologiques et techniques posés par le principe de l'information en n-dimensions, les cartographies que j'ai produites jusqu'à présent paraissent toutes relatives. Ce ne sont effectivement que quelques clichés photographiques d'une première apparence derrière laquelle il faut désormais deviner toutes les autres dimensions. Ainsi, le long travail de documentation d'un phénomène donné, surtout s'il est social, suppose la mise en œuvre d'un modèle de données plus complexe, apte à rendre compte des différents pans de réalité observée.

1. European FET project MULTIPLEX, "Foundational research on multilevel complex networks and systems" (Reference Number 317532), <http://arxiv.org/pdf/1407.0742v2.pdf>.

Il est temps d'examiner les trois implications majeures de ces nouvelles orientations de la recherche technologique sur les réseaux qui impactent fortement la pratique de la cartographie d'information (fût-elle adossée à un projet générique de cartographie du Web). La première ressemble à un défi lancé aux ingénieurs et aux chercheurs en ingénierie de l'information, mais aussi aux mathématiciens et à tous ceux qui conçoivent et testent des algorithmes : les modèles de données multidimensionnelles peuvent (ou doivent) être l'objet d'opérations plus complexes que l'addition-soustraction.

Si l'on revient à l'étude de Valdis Krebs, les fils qui ont conditionné le destin des kamikazes du 11 septembre s'appuient sur l'exploitation de 4 sources différentes (*Trust, Task, Money and Ressources, Strategy and Goals*) que le chercheur finit par agréger en *une seule* structure de graphe pour pouvoir la traiter avec les instruments dont il dispose. Cependant, qu'en est-il si l'on considère les quatre types de connexion comme des graphes indépendants qui interagiraient de différentes façons : quels en sont les nœuds communs ? Sont-ils hiérarchisables ? Les connexions des quatre types sont-elles simultanées ou consécutives ? Certaines propriétés de structure dans un *layer* en impliquent-elles d'autres dans un autre *layer* ? Dès le départ, depuis la construction de la table de données, les informations agrégées par Valdis Krebs sont donc réellement *multidimensionnelles* et, potentiellement, elles méritent d'être traitées comme des *variables* dont les corrélations (et pas seulement l'agrégation) devraient être interrogées.

C'est ici que l'on aperçoit les limites de la boîte à outils habituelle pour comprendre comment deux plans ou dimensions de la réalité (sociale en ce qui concerne les organisations criminelles) entrent en interaction. Ainsi, plus on dispose de colonnes dans un fichier Excel, d'attributs pour décrire les nœuds d'un graphe, de types de liens dans un réseau, plus se pose la question de savoir comment les corréler. À supposer que l'on arrive à isoler des formes typiques d'interaction entre les liens de type 1 et

les liens de type 2, qu'en est-il alors avec les liens de type 3, de type 4, etc. ? Les structures de graphes multidimensionnels présentent donc un potentiel explicatif puissant, mais à partir de calculs complexes tels qu'ils sont encore largement hors de notre portée. La *corrélation de dimensions* dans ce que l'on appelle « information » n'est donc pas à l'évidence l'*addition* : elle suppose fondamentalement l'établissement de relations logiques ou causales entre deux phénomènes (pour nous, deux *patterns* ou certaines de leurs propriétés) situés dans deux dimensions différentes, éventuellement deux réseaux distincts qui interagiraient. Dans une telle perspective, une entité (sociale ou non) doit être considérée comme participant à une pluralité de types d'interaction ou de connexion avec les autres. Ainsi, peut-on établir un ou des liens entre un phénomène observé dans une dimension (par exemple une configuration typique des réseaux de croyances et de pratiques partagées) et un autre situé dans une autre dimension (par exemple l'activation d'un réseau de soutien logistique) ? Et apparaissent-ils de façon simultanée, de façon consécutive ou bien sur des périodes de temps encore jamais prises en compte (des années, une vie) ? Là, il ne s'agit plus d'observer un aspect du social avec les anciennes lunettes qui scindaient en deux nos univers entre le *micro* et la *macro* (et tous les niveaux de *mezzo* que l'on voudra bien imaginer), mais de la réalité multidimensionnelle du plus singulier des objets aux plus grandes masses. En ce sens, les graphes à produire désormais comme les solutions cartographiques associées doivent pouvoir supporter de l'information en n-dimensions, qu'il s'agisse de description, de prospective, voire de prédiction pour s'en tenir aux questions de sécurité. La maîtrise des graphes *multilayers* serait d'une grande conséquence pour nos capacités à observer des organisations ou des mouvances terroristes et djihadistes. Par exemple, quels sont les rapports possibles entre les *patterns* exhibés par les liens familiaux et ceux exhibés par les relations nouées autour de croyances religieuses ? Doit-on l'existence actuelle de certaines relations à la réactivation de plus anciennes, issues d'un autre plan de réalité ? Quelle variable, ou association de variables, paraît la plus pertinente pour expliquer le projet d'attentat ? Et de façon complémentaire, quelles variables de connexion (ou *layers* de liens) jouent un rôle majeur dans l'organisation logistique des attentats ?

Évidemment, je semble ainsi découvrir naïvement dans le champ de la cartographie ce que d'autres ont appris à maîtriser depuis longtemps, notamment en économétrie, en statistiques et en mathématiques appliquées sous le nom de « système de multivariées ». Les bases des différentes méthodes de recherche en composantes principales sont en effet posées depuis longtemps (avec des applications par exemple en sémantique), mais ne semblent guère utiles dans une approche par graphe : de la plus importante à la plus modeste variable dans un jeu de données, il s'agit de comprendre la combinatoire qui donne naissance à des topologies particulières avec leurs propriétés de structures (centralité, *clustering*, hiérarchies...). Si les statistiques

peuvent aider, les propriétés sous-jacentes liées aux espaces relationnels restent encore à découvrir, en espérant que l'on pourra les mathématiser. Considérer l'information comme objet multidimensionnel ouvre des perspectives fascinantes en termes de recherche sur les graphes et d'expérimentations sur des jeux de données réels, mais présente de nombreux verrous méthodologiques et technologiques. Clairement, il s'agit pour moi, à titre de cartographe, de nouvelles frontières scientifiques. On peut ainsi rêver à une version future de Gephi qui, dans une dizaine d'années ou peut-être plus, embarquera un HITS ou un PageRank dédié à l'exploration de propriétés de graphes sur plusieurs échelles. Ils nous diraient, par exemple, qu'étant donné tel type de graphe ou de structure, si A est vérifié dans un *layer* 1, alors B se vérifiera dans un *layer* 2 avec tel effet sur un *layer* 3. Et ainsi de suite... selon une immensité de règles combinatoires possibles dont certaines seulement seront sélectionnées pour leur pouvoir explicatif sur des jeux de données réels. Et peut-être aussi pour leur pouvoir prédictif.

La seconde implication majeure de la multidimensionnalité des données concerne les interfaces d'accès et, au-delà, les scénarios d'usages associés. Il s'agit d'un point central pour le cartographe que je suis puisque la question de l'interface finale (disons « centrée utilisateur ») inclut de fait aussi bien les techniques de spatialisation de graphes que les scénarios d'usage (ou les « contextes métier ») qui accompagneraient les futurs explorateurs de *data*, et donc le *design* global d'un futur système. L'univers de la cartographie des organisations criminelles ou terroristes constitue, là aussi, un terrain précieux d'expérimentations et riche d'enseignements. Depuis l'étude sur les attentats du 11 septembre 2001 réalisée à partir des données de Valdis Krebs, j'ai relancé il y a quelque temps le projet de construction de graphes permettant d'appréhender les relations multiples et complexes entre les membres de cellules opérationnelles, mais aussi en incluant toutes les entités (individus ou organisations) qui seraient liées à des attentats issus de la mouvance islamiste. Avec l'aide de deux jeunes ingénieurs (Céline X et Léo Y), nous avons repris les données de l'étude de Valdis Krebs et intégré deux autres jeux de données pour faire émerger de nouvelles pistes à explorer en termes de représentation et de spatialisation de graphes. Le premier a été construit en décembre 2005 par José Rodríguez lorsqu'il a produit son étude sur l'organisation des attentats à la gare d'Atocha en banlieue de Madrid en 2004². Les données recueillies par le chercheur concernent 70 acteurs

2. J. A. Rodríguez, "The March 11th Terrorist Network: In its weakness lies its strength", 2005, <http://citeseeerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.98.4408>.

(dont les 7 terroristes) reliés entre eux par 5 types de liens. Les 5 *layers* de liens permettent d'identifier différents types de connexion (liens d'affinités entre les 7 acteurs principaux, mais aussi liens d'amitié, liens de proximité pour tous ceux qui se trouvent dans la région de Madrid, liens de confiance à partir de critères religieux et liens historiques produits lors de séjours en Afghanistan et au Pakistan).

Image 10.5

Réseau de José Rodríguez et ses différentes couches.

Source : J.A. Rodríguez, "The March 11th Terrorist Network: In its weakness lies its strength", 2005, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.98.4408>.

Ainsi, le graphe complet de la vue de gauche peut être décomposé en 5 *layers* que l'on peut superposer de différentes façons (dans la partie droite), comme l'ont réalisé Céline et Léo à partir de feuilles transparentes. Le second jeu de données a été repris et transformé en structure de graphes analysable avec Gephi : il s'agit des

informations collectées par Sean Everton et ses étudiants concernant les acteurs et les organisations liés aux attentats de 2000 à 2006 en Indonésie, notamment à Bali³. Le réseau Noordin Mohammed Top (du nom du *leader* du groupe) comprend des informations sur 80 acteurs distribués en 13 *layers* différents. Pour visualiser les données et la richesse informationnelle des 13 *layers*, Léo et Céline ont imaginé une interface graphique pouvant être intégrée à un futur prototype dédié à ce type de jeux de données :

Image 10.6

Processus de construction d'un graphe *multilayers*.

L'interface de visualisation est issue d'une procédure de transformation des tableaux de données (1) en structure de graphe *multilayers* (2), elle-même spatialisée (3) selon deux règles. La première consiste à placer dans un cercle central tous les acteurs identifiés ayant *directement* participé aux attentats (on peut donc les intégrer au fur et à mesure d'une enquête). La seconde consiste à positionner à l'extérieur du cercle les acteurs ayant participé *indirectement* aux attentats en les positionnant selon des règles « classiques » d'attraction/répulsion (par exemple avec le Force Atlas 2 avec Gephi).

3. N. Roberts et S.F. Everton, "Strategies for Combating Dark Networks", *Journal of Social Structure*, 12 (1), 2011, https://www.exeley.com/journal_of_social_structure/doi/10.21307/joss-2019-030.

Image 10.7

Principe de représentation d'un réseau *multilayers*.

Dans le système imaginé, les *layers* de liens peuvent être affichés ou non selon des réglages contrôlés par un analyste ou un expert (ici, tous les *layers* ont été affichés). De plus, il permet de décliner la variable visuelle des couleurs en fonction des différents attributs présents dans le jeu de données. Dans le cas de la campagne d'attentats menés en Indonésie de 2000 à 2006, les couleurs ont été attribuées en fonction de l'appartenance des acteurs à des organisations actives, comme la Jemaah Islamiyah qui est la plus présente (certaines autres organisations présentes dans le graphe y étant liées par ailleurs). Les choix graphiques de spatialisation tiennent évidemment de présupposés ou d'hypothèses de travail ; comme le montre la vue générale, il s'agit pour nous d'essayer de comprendre comment un ensemble d'organisations islamistes (ou d'individus y appartenant) peuvent contribuer à alimenter des cellules actives (le réseau Nordin Mohammed Top est donc, en quelque sorte, *multiorganisationnel*). Il serait intéressant, sur ce point, de générer l'espace global de l'islamisme radical sunnite (par exemple les cellules issues du wahhabisme) où l'on verrait apparaître successivement des zones de formation de cellules actives au cours du temps. Y figureraient, assurément, les acteurs décrits dans nos trois jeux de données...

Une troisième implication (et non la dernière) du principe de la multidimensionnalité de l'information concerne la révision qui m'apparaît nécessaire de nos outils, de nos méthodes... et de nos concepts théoriques. Là aussi, la cartographie des organisations criminelles et terroristes apparaît singulièrement éclairante. Ainsi, si l'on revient à la technique de superposition de liens « faibles » et « forts » (ou selon toute échelle de pondération plus complexe) pour analyser différents types de connexions connues qui animent un espace social (et dont la logique ou la dynamique fait l'organisation), il faut alors prendre en compte les différentes formes ou topologies de chacun des réseaux, et donc leurs dynamiques temporelles spécifiques. Ainsi, dans un graphe *multilayers* ou de différents types de connexions, A est une entité qui interagit avec d'autres entités d'un certain point de vue (par exemple des flux bancaires) tandis que B interagit parallèlement avec un autre groupe (via des *e-mails*). Et, tout d'un coup, A appelle B via son téléphone mobile, dressant ainsi un pont (principe du *bridging*) entre A et sa communauté et B et sa communauté parce que A et B se sont entraînés ensemble en Afghanistan. Ce que l'on présente en sociologie comme des liens faibles dans les sociographes (ces grands liens qui permettent à un individu de « sauter » d'un groupe à un autre en termes de parcours) se révèlent à l'analyse comme d'anciens liens forts, endormis dans l'histoire de chacun. L'étude de Valdis Krebs comme notre modeste cartographie exhibent une série de liens faibles (connexions ponctuelles de type communication électronique) qui peuvent être considérés, à bien des égards, comme la « réactivation » d'anciens liens forts. Dans le cas du 11 septembre 2001, la dynamique des cellules opérationnelles (liens de coparticipation dans un commando, liens de soutiens logistiques, toutes les actions de préparation des attentats...) repose largement sur la réactivation de liens forts plus anciens, longtemps restés en sommeil et qui se réveillent à différents niveaux du réseau social.

Dans la tradition de la modélisation statistique (par exemple chez Watts et Strogatz), les liens faibles sont distribués aléatoirement (principe du *randomness* assumé pour la démonstration statistique) de façon à démontrer le principe du *small world* où quelques grands ponts (*bridging*) sont jetés d'un bout à l'autre du graphe faisant ainsi chuter le diamètre d'une structure au départ très clusterisée (en chemins réels ou possibles). On connaît aussi leur importance chez Stanley Milgram et les « chain of acquaintances » (1969) et, surtout, Mark Granovetter (1973). Chez ce dernier, les *weak ties* tirent leur force dans l'étude de la macrostructure sociale de leur pouvoir de connexion de différents univers sociaux : tandis que les *strong-local ties* permettent de décrire les clans familiaux et les univers d'amis proches, les *weak-global ties* semblent jouer un rôle déterminant dans notre faculté d'adaptation sociétale au sens large (ils sont activés dans des périodes de recherche d'emploi chez les jeunes diplômés). Mais rien ne nous dit dans quelle(s) proportion(s), ni selon quelle(s) formule(s) d'intrication ces liens faibles ne sont pas en même temps d'anciens liens

forts de proximité sociale. À l'examen, il se pourrait donc bien que les liens faibles n'existent pas, du moins leur principe théorique tient tant que l'on n'a pas engrangé suffisamment de données sur la « surface » d'un phénomène réseau ou questionné l'histoire de ses connexions, des plus anciennes au plus actuelles. On retrouve là l'un des problèmes majeurs des SHS aujourd'hui : faute d'instruments performants pour interroger les données (en précision qualitative comme en largeur quantitative), les chercheurs (dont les sociologues) sont restreints à des horizons d'observation relativement réduits.

Et ce n'est là qu'un tout premier aspect de la revisitation des méthodes et des concepts que nous livrent les sciences (humaines ou non) à la lumière des problématiques de la multidimensionnalité de l'information. Là encore, les questions de sécurité et de l'observation du djihadisme sous l'angle des réseaux posent d'intéressants problèmes qui ne sont pas que technologiques ou méthodologiques : en braquant l'attention sur une « zone » particulière du social où des entités interagissent selon des scénarios ou des *patterns* (que l'on espère pouvoir observer, puis analyser, classer, comparer), à partir de quels concepts, voire de quelle théorie, appréhender la dynamique propre à chaque *layer* de liens, sans parler de leurs interactions multiples ? Tout m'avertit que là où la question de la multidimensionnalité de l'information fait surface, comme le domaine de la sécurité, se forgeront demain les méthodes et peut-être les concepts qui éclaireront d'autres systèmes, comme le Web et sa structure fascinante. Les univers informationnels auxquels il faut se préparer ne seront pas réductibles à des types, des classes ou profils et auront tout d'instruments d'observation de la dynamique complexe des interactions et de leurs dimensions. À ce titre, il n'est pas inutile de considérer, là aussi, le djihadisme actuel comme un écosystème social et culturel où les *patterns* de connexion sont complexes comme le montrent les données des attentats du 11 septembre, puis ceux de Madrid en 2004 et en Indonésie jusqu'en 2006 : diversité des profils et des parcours personnels, groupement en *clusters* (cellules, groupes, filières), différentes formes de hiérarchies (les plus actifs, ceux qui jouent un rôle central dans la structure, les mentors qui diffusent le message, recruteur dans les filières, multiples formes de soutien...), capacités de résilience organisationnelle. Cet écosystème semble fonctionner sur plusieurs niveaux (*layers*) que savent exploiter les acteurs qui passent de l'un à l'autre, qui utilisent l'un pour l'autre. Les entités sociales (individus ou collectifs) se caractérisent par une sorte de « plasticité » ou de « mobilité » qui dessine une nouvelle géographie où se croisent les trajectoires personnelles dans des cellules de rencontres (les cybercafés, les mosquées, les librairies), dans des villes nœuds et dans des bases d'entraînement, dans les nouvelles terres du djihad telles que l'Afghanistan et l'Irak. De façon simultanée, le Web et ses plateformes sociales (parfois cryptées) se constituent en instruments structurant l'idéologie djihadiste ou en dispositif

logistique d'échanges d'informations. Actifs ou en sommeil, les réseaux djihadistes d'aujourd'hui relèvent d'un espace tout aussi réel que virtuel : les recruteurs d'un côté, le Web de l'autre s'étendent de plus en plus jusqu'à eux, pour les inclure dans une « nation virtuelle » dont ils peuvent devenir des soldats. Voilà un objectif pour un cartographe ! Et quelques aventures supplémentaires...

Chapitre II. Une première plongée dans le réseau Tor

Le journal *Le Monde* a publié en juin 2017 un article sur les dernières recherches consacrées au fameux « *dark web* », et plus particulièrement le réseau Tor (The Onion Router). L'article fait référence à une étude que nous avons menée régulièrement depuis quelques mois (F. Ghitalla, F. Boisselier, J. Decayeux, E. Bartholmé, V. Déhaye), en récoltant petit à petit une série d'indices sur le fonctionnement du réseau Tor pour en comprendre la nature... et certains de ses enjeux.

Image 11.1

Réseau de l'étude de Tor présentée dans *Le Monde*. Répartition des 5 043 nœuds du *dark web* et de leurs 7 275 liens. Les couleurs correspondent à des agrégats de sites très liés entre eux, mais peu avec les autres. F. Ghitalla, F. Boisselier, J. Decayeux, E. Bartholmé, V. Déhaye.

La cartographie publiée, qui peut faire sensation, ne révèle pourtant qu'une partie infime de ce réseau obscur. Examinée dans le détail, on peut en tirer quelques enseignements utiles pour la compréhension de la nature de son fonctionnement (une question technique qui n'est pas triviale pour un néophyte comme moi), mais aussi pour la place qu'occupe Tor (ou d'autres réseaux cryptés et décentralisés) dans l'écosystème global du net.

Cette étude permet d'entrevoir quelques-unes des propriétés fondamentales de ce *dark web* que j'aime à représenter, selon la vision commune, sous la forme d'un vaste continent enfoui sous les profondeurs comme je l'ai fait dans le présent ouvrage.

Cependant, on aurait tort de se le représenter comme un espace figé, presque minéralisé, tellement secret ou enfoui que seuls quelques initiés pourraient atteindre. Cette première étude montre qu'il est largement ouvert à tous, mais aussi animé de mouvements très dynamiques, si l'on pense, par exemple, à toutes les ressources et les arguments qui y ont cours sur les questions de l'anonymat, des nécessités de se protéger contre les instruments de traçabilité de l'humain en ligne ou de la liberté d'informer. D'une certaine façon, le *dark market* auquel on réduit trop souvent le réseau Tor, et les technologies type *bitcoin* qui se profilent au-delà, contribuent à une dynamique d'innovation indéniable.

Personnellement, je verrais Tor comme un refuge, une sorte de sanctuaire où viendraient prendre place tous ceux qui ont de bonnes et de mauvaises raisons de vivre cachés. Pour comprendre Tor, ou plus généralement le *dark web*, il ne faut pas le confondre avec la strate du *deep web* (autrefois désigné comme *invisible*) où sédimementent des masses documentaires qui m'apparaissent plus « statiques » (informations légales, documents techniques, certaines données personnelles, bases de connaissances, dossiers médicaux...), souvent accessibles derrière un identifiant et un mot de passe depuis le Web de surface (*clear web*). Tor est une sorte de toile (vraiment) décentralisée d'ordinateurs dotés d'un cryptage, un réseau privé, mais, on le verra, très facile d'accès. Avant cette étude, Tor était pour moi associé à deux aspects contradictoires de ce Web crypté et donc anonyme : « l'affaire » Snowden en 2013 et les *dark nets* ou *dark markets* où tout, ou presque, peut s'acheter ou se vendre, depuis des numéros de carte bleue jusqu'à la drogue en passant par... tout ce que l'on peut envisager de vendre de façon illégale. Rien de surprenant à ce que la justice américaine s'y intéresse de près (l'affaire Silk Road date aussi de 2013) tout comme la NSA ou d'autres agences gouvernementales. Mais Tor semble résister, et l'on ne peut être que surpris par l'activité des « silos » que nous avons visité en ce printemps 2017, en passant de *hidden service* en *hidden service* dans une navigation sans un Google général dédié au domaine qui en recenserait tous les aspects.

L'équipement

Pour visiter Tor, il faut être équipé, car les sites en *.onion* du réseau crypté sont accessibles via une série de couches successives sous lesquelles se cache le trafic « réel ». C'est la métaphore de l'oignon ou le routage en oignon développé dès les

années 1990 par le laboratoire de mathématiques de l'US Navy, puis par la DARPA. Mais c'est en 2001 que la version alpha de Tor est lancée et, depuis, le *dark web* se développe à l'abri du Web de surface ou *public web* et dont des acteurs (souvent bénévoles) aménagent de nombreux passages ou accès sous forme de tunnels, de plus en plus nombreux et accessibles. L'équipement minimal pour s'y plonger consiste à s'équiper d'un *navigateur Tor* (spécialement dédié au traitement des adresses en *.onion*) et d'un VPN. Un VPN ou Virtual Private Network permet de générer depuis votre PC des accès au Web via un autre ordinateur présent sur le réseau. Le protocole de communication crypté Tor et un VPN constituent une forme d'équipement minimal assez facile à trouver : un VPN, par exemple sur Top VPN software un navigateur Tor sur Tor2web et on peut commencer par explorer les liens proposés par Darkweb News.

Image 11.2

Exemple d'un point d'entrée vers le *dark web*.

Le fait de s'équiper ainsi peut conduire certaines agences gouvernementales à considérer l'explorateur comme un suspect potentiel. C'est l'un des aspects des révélations d'Edward Snowden, mais, comme nous l'apprend Kevin Mitnick dans son dernier ouvrage *The Art of Invisibility*, il semble presque impossible d'échapper à toute forme de contrôle (bienveillant ou malveillant). Les points de passage du *clear web* à Tor semblent particulièrement surveillés, à l'entrée (*entry guard*) du réseau comme à la sortie (*exit node*). On trouvera dans *Wired* en 2014 un article intéressant résumant quelques-uns des aspects essentiels de cette surveillance des usages sur les réseaux, "Use privacy services? The NSA is probably tracking you".

Une fois plongée dans l'univers Tor, la navigation peut s'avérer lente, car les communications obéissent à une série de principes techniques particuliers qui rendent le réseau imperméable à l'observation détaillée de l'activité de chacun des usagers. En effet, les connexions sur Tor s'effectuent :

1. à travers au moins 3 *routers* (*nodes*), sauf en entrée et en sortie ;
2. sélectionnés au hasard (*random pathway*) parmi des milliers de *routers* possibles (maintenus par des milliers de volontaires dans le monde) ;
3. en transmettant une information cryptée (le client et le serveur étant dotés des clefs cryptographiques de chacun des *routers* utilisés) (AES – Advanced Encryption Standard) ;
4. divisée en plusieurs parts égales (*cells*) ;
5. à travers le *pipe* défini (tuyau construit entre les *routers* pour la transaction/communication) où chaque router ne connaît que le précédent et le suivant ;
6. l'information étant finalement reconstruite via un *router* de sortie avant de l'envoyer vers le client final. C'est tout le principe du réseau Tor qui repose donc sur une combinaison hybridant cryptographie et routage des communications par de multiples chemins potentiels. Comme des pelures, les différentes clefs de cryptage opèrent à chaque étape du processus :

Image 11.3

Structure du réseau Tor.

En termes de topologie réseau, on peut en résumer le principe de la façon suivante :

Image 11.4

Structure du réseau Tor.

À l'échelle mondiale, on imagine ainsi l'intensité du trafic entre serveurs Tor si chaque connexion client-serveur génère à chaque fois des chemins aussi longs de façon dynamique, et pour des communications cryptées à chaque étape. C'est pourquoi le programme de la NSA de surveillance (du moins, tel qu'il est présenté depuis les révélations d'Edward Snowden avec l'utilisation, notamment, de XKeyScore) se concentre sur tout ce qui est techniquement accessible en termes de traces d'usage : les internautes qui s'équipent (en téléchargeant un navigateur Tor ou un VPN), ceux qui accèdent à des sites ou des serveurs réputés « alternatifs » comme Linux Journal, les utilisateurs de ressources pour installer des outils de cryptage ou de gestion de l'anonymat comme Tails, HotSpotShield, Freenet, Centurion, FreeProxy, Megaproxy...

Les chemins

Après avoir visité quelques sites du réseau Tor, on se rend vite compte que la navigation semble à chaque fois très localisée et organisée « en silos » : à partir de quelques listes d'adresses, souvent classées en thèmes (drogue, matériel électronique, fausse monnaie... dans les « zones du commerce »), on peut commencer à explorer ce qui nous est apparu massivement comme une série de marchés, ou plutôt de bazars où les biens et les services s'entassent sans ordre apparent et de façon cloisonnée... et dont les adresses changent rapidement. Mais d'autres chemins possibles se mêlent aux accès aux *hidden services* de ce bazar secret, y compris ceux qui nous font « remonter » vers le Web public que l'on peut alors arpenter de façon anonyme.

C'est ainsi que notre petit groupe d'explorateurs a eu l'idée d'essayer de recenser une partie significative de tous ces chemins de façon automatique en « crawling » les ressources accessibles. Ainsi, en quelque 3 semaines, nous avons obtenu un fichier de graphe comportant 8 328 sites (et parfois simplement des pages uniques) reliés entre eux par plus de 13 000 liens. L'idée que nous poursuivions était de recenser une partie des sites en *.onion* et leurs connexions mutuelles, mais aussi d'intégrer au corpus des sites ou des pages du Web public en *.com*, *.fr*, *.edu...*) qui permettent d'accéder au réseau Tor. Nous espérions trouver, notamment, ces fameux « points d'entrée » (*bridges*) particuliers que cherchait à surveiller la NSA et qui assurent le passage d'un Web à l'autre.

Image 11.5

Étude : extraction (*crawling*) d'un corpus composé de 8 238 sites web reliés entre eux par 13 277 liens dont 5 043 *hidden services* du réseau Tor (61 % du corpus) en rouge, et 3 195 Clearnet ou Web public (39% du corpus). Les ponts (*bridges*) entre le Web public (*clear web*) et les sites *.onion* du réseau Tor sont nombreux et variés. Les « deux Web » sont largement imbriqués l'un dans l'autre.

Dès les premières analyses avec Gephi, nous avons été surpris par le foisonnement des chemins possibles entre les *clear web* (en bleu) et Tor (en rouge). On peut passer massivement de l'un à l'autre et dans les deux sens (notre *crawler* étant compatible avec les deux univers). Une autre vue produite avec Gephi fait bien apparaître l'intensité des liaisons quasi organiques entre les deux mondes.

Image 11.6

Intenses liaisons entre le *clear web* (en bleu) et le *dark web* (Tor, en rouge).

Peut-on d'ailleurs encore parler de « deux mondes » à partir de cette première expérience ? Pas si sûr, car en 2017 le Web n'est probablement plus le même qu'au moment de l'affaire Snowden : Tor semble s'être popularisé au point d'être hybridé avec le Web public, même si techniquement les zones d'anonymat dessinées par le protocole Tor restent parfaitement distinctes.

Mais d'où viennent toutes les pages qui font office de tunnel du *clear* au *dark web* ? Sont-ce des blogs de cyberpunks, des forums d'activistes, de crypto-anarchistes ou de *hackers* ? Certaines clairement, mais la majorité apparaît d'une grande variété de « communautés » dans l'écosystème du Web. Pour le vérifier, nous avons analysé manuellement 295 URL du *clear web* ayant plus de 3 liens pour rendre compte de cette diversité. Nous les avons classés en grands *types* ou *classes*, même si certaines adresses relèvent de plusieurs catégories. En essayant d'aligner nos classes, on s'aperçoit rapidement du lent dégradé qui nous fait passer des sites web les plus populaires aux sites plus techniques ou les plus spécialisés :

Image 11.7

Principaux types de sites sur Tor.

Sans être un spécialiste de toutes les subtilités techniques de certaines de ces ressources du Web public, on peut ainsi deviner quelques grands ensembles de classes. Le premier d'entre eux est constitué de pages « classiques » issues des réseaux sociaux (Facebook, Twitter, LinkedIn...), des plateformes de blog (type WordPress) ou des pages ramenées depuis les moteurs de recherche (15,25 % du corpus). On peut intégrer à ce premier ensemble des sites d'informations générales (type presse et médias en ligne soit 8,47 % du corpus). Les deux classes sont en bleu dans notre échelle.

Vient ensuite un ensemble plus vaste, et plus spécialisé aussi d'un point de vue technique et informatique composé de sites dédiés aux technologies de l'information orientées réseau et Web, ceux animés par les communautés *open source* et des sites plus engagés dans la défense de l'anonymat sur le Web. On peut ranger dans le premier groupe des sites comme <http://www.privoxy.org/>, <https://retros-hareteam.wordpress.com/>, <https://www.cryptoparty.in/> qui constituent 11,52 % du corpus dans le second groupe dédié à l'*open source* des sites comme <https://tails.boum.org/>, <https://whispersystems.org/>, <http://www.pidgin.im/>, <http://www.simbionica.net/>... (constituant 25,08 % du corpus). Les deux ensembles se confondent presque et semblent s'associer logiquement au troisième dédié aux sites autant militants que technophiles que sont par exemple <https://www.guerrillamail.com/>, <https://www.piratenpartei.at/>, <https://www.internetdefenseleague.org/>, (soit 22,37 % du corpus).

Enfin, un dernier ensemble d'URL semble se constituer autour de deux projets technologiques complémentaires, le réseau Tor lui-même et l'univers du *bitcoin* et de la *blockchain*. S'il n'est pas aisé de s'y plonger sans compétences techniques avancées, il apparaît manifeste que les URL classées dans ces deux ensembles sont alimentées par des acteurs et des communautés très actives et très inventives, des foyers d'innovations technologiques en mouvement permanent comme <https://blog.cryptographyengineering.com/>, <https://blockchain.info/>, <https://keybase.io/>, <http://dogechain.info/> dans le domaine de la crypto-ingénierie (9,15 % du corpus) ou les sites du projet Tor comme www.torproject.org, <https://>

www.freehaven.net/, <http://torstatus.blutmagie.de/>, <https://atlas.torproject.org/>, <https://db.torproject.org/>... (8,13% du corpus).

Tor et son *dark market*

Il faudrait pousser plus loin l'analyse des rapports entre public et *dark web*, en particulier en termes de types de distribution des liens ou des connexions (pour déterminer, notamment, quels types de sites web publics pointent vers quels types de sites en *.onion*). Néanmoins, notre *crawl* sur Tor montre combien les différents marchés sont présents dans les ressources identifiées (avec toutes les précautions que suppose un *crawl* en termes de méthodologie et de résultats). Le *dark market* semble en effet particulièrement actif, et désormais suffisamment attractif en 2017 puisque Google en indexe une bonne part. Il suffit pour cela de consulter Nucleusmarket, Alphabay-market ou le « top 50 » des *.onion* produit par Deepweb-sites, tous indexés par nos moteurs de recherche préférés.

Notre exploration automatique (*crawl*) des liens a été l'occasion d'indexer les contenus, en particulier les mots ou expressions les plus présents dans plus de nos 5 000 références en *.onion*. Pour le moment, cette étude statistique des contenus se limite à un graphe de co-occurrences de termes (un terme A est lié à un terme B quand ils apparaissent dans la même URL) permettant de classer les termes en fonction de leur intensité dans le corpus et de les associer entre eux. L'examen des termes indexés montre la présence massive (dans le cadre de notre corpus) des différents *hidden services* du *dark market*. Le vocabulaire le plus massivement utilisé concerne les questions de sécurisation des transactions et la protection de l'anonymat.

Image 11.8

Réseau de co-occurrences des termes utilisés par les sites du *dark market*.

Le graphe de co-occurrences laisse apercevoir une série de *clusters* reliés entre eux, essentiellement constitués par un vocabulaire technique. Plus de la moitié des termes présents dans la base ayant un degré d'occurrence égal ou supérieur à 5 font référence à l'infrastructure de vente et d'achat sur le *dark market* (en couleurs dans les vues produites avec Gephi). Nous avons ensuite analysé chacun des *clusters* identifiés (avec l'algorithme Modularity et différents réglages), qui font nettement apparaître deux ensembles complémentaires : en rouge, le vocabulaire associé à la gestion de l'infrastructure des transactions et en jaune, la gestion de l'infrastructure de paiement.

Image 11.9

Mots-clés le plus souvent utilisés se référant à l'infrastructure de vente dans le corpus / le processus de vente.

Une telle concentration de vocabulaire technique n'est pas étonnante pour ceux qui ont l'habitude de naviguer sur Tor : les sites accessibles sur ce réseau sont relativement « statiques », faisant penser au Web 1.0 des premiers temps. Mais, surtout, de nombreux sites marchands proposent des biens sous forme de catalogues d'images où figure peu de texte. Le texte accompagne surtout les explications techniques, les opérations de cryptage et les procédures à suivre pour les acheteurs. C'est pourquoi, dans la seconde moitié du graphe de co-occurrences, nous avons figuré en vert le vocabulaire lié essentiellement aux opérations de contact et de procédures d'achat côté utilisateur ou côté client. Le *cluster* qui figure en bas du graphe est exemplaire de « qualité » du service et des garanties apportées aux acheteurs à un niveau mondial (comme tout achat légal en ligne !).

Image 11.10

Mots-clés le plus souvent utilisés se référant à l'infrastructure de vente dans le corpus / le processus d'achat client.

Côté « marchandises », une première analyse des « biens et services » proposés illustre bien la réputation du réseau Tor : on y trouve, comme l'indique un site de référence du *dark net* :

Most of these Darknet Markets offer things like drugs, counterfeit money and IDs, stolen credit cards, weapons, all kind of accounts for porn and adult websites, lifetime Netflix subscription, stolen ancient artifacts ; you can buy services from hackers.

Nous avons repéré manuellement, parmi les mots-clés et les expressions le plus souvent présents dans notre base, ceux qui évoquent ce marché et l'étendue des biens ou services disponibles :

Image 11.11

Sélection des termes les plus représentatifs.

Si la nature des biens vendus peut surprendre, il apparaît que le *dark market* est très divers et qu'il n'est pas réductible à un marché de la pédophilie comme l'a montré récemment un article de *Wired* en réponse au département de la justice américaine. Toujours est-il que, malgré les spécificités techniques de Tor et la nécessité de s'équiper spécifiquement pour y accéder, on ne peut qu'être frappé par l'accompagnement ou la « prise en main » du client. Comme dans tout marché, différents modes de paiement sont possibles, essentiellement avec des crypto-monnaies, bitcoin évidemment, mais aussi parfois Litecoin ou Darkcoin... Et comment s'en procurer ? Rien de plus simple puisque le Web public « de surface » propose différents types de transaction pour en obtenir comme Paxful.

Un continent d'innovations

L'organisation des ressources « en silos » séparés autorise évidemment que coexistent les *hidden services* et d'innombrables ressources concernant la protection des données personnelles et de la vie privée et, au-delà, l'expression d'opinions politiques et idéologiques alternatives. Ces espaces d'expression libre sont en général associés à un discours sur les technologies de cryptographie, souvent accompagnés de *guide-lines* sur les mesures usuelles à adopter par chacun depuis sa propre machine.

Sur plus de nos 8 000 URL indexées au cours du *crawl*, plus de 3 000 sont issues du Web public et assurent des chemins vers le Tor et le *dark web* ou l'inverse. Certaines de ces 3 000 URL ne conduisent pas vers le *dark market*, mais vers différents « lieux » (a priori assez circonscrits) que l'on peut qualifier d'alternatifs, où se mêlent arguments politiques et défense de l'anonymat, voire de l'action collective anonyme. Pour le moment, il est encore trop tôt pour en proposer un classement en termes de contenus, ou même d'en dresser une carte suffisamment exhaustive. Néanmoins, on s'aperçoit de la vitalité de ces différents lieux en termes de participation, de diffusion de la technologie Tor et de ses usages « citoyens ». On trouve ainsi dans notre corpus des références au projet ProPublica, qui est annoncé comme le premier média d'information du réseau Tor. Le succès élargi du réseau Tor dans le domaine du journalisme depuis WikiLeaks s'illustre aussi dans une formation (que je ne connaissais pas) du Centre de formation du journalisme (CFJ) sur le *dark web* et ses usages dans le domaine de l'information. On trouve quelques références à certaines initiatives de plus grande ampleur comme le partenariat entre Reporters sans frontières et Torserver.net dans la lutte contre la surveillance et la censure en ligne.

On ne peut guère entamer un (court) voyage dans ce *dark web* militant sans rencontrer la communauté Anonymous en particulier à travers www.anonymous-france.info qui indique (notamment) comment soumettre des documents à WikiLeaks pour analyse. Comme l'indique la page : « WikiLeaks propose aux lanceurs d'alertes, activistes, hackers, journalistes et citoyens, une plate-forme sécurisée pour envoyer des documents confidentiels. Le système de soumission de documents associe la protection et l'anonymat grâce à l'utilisation du réseau Tor "The Onion Roteur" (pour le Roteur Oignon). Il vous faudra impérativement installer le "Tor Browser Bundle" (TBB) pour pouvoir y accéder ».

Mais si des chemins sur Tor nous mènent vers des zones d'innovation sociale et citoyenne, d'autres nous ont aussi amenés vers des lieux d'innovation technologique continue et particulièrement actifs comme Bitcoin et la *blockchain*. Personnellement, j'en suis encore à m'initier à ces domaines qui sonnent comme des mots magiques

depuis peu. La *blockchain* (Bitcoin n'étant que l'un de ses aspects), c'est d'abord un projet fascinant soutenu par des communautés d'acteurs, à mon sens la plupart du temps des experts de haut niveau qui travaillent sur les mécanismes de stockage et de transmission d'actifs immatériels transparents, sécurisés et sans intermédiaire. Si vous allez sur les sites web publics comme Chaintech, Asseth, le Cercle du coin, ou Ethereum, vous serez peut-être surpris comme moi par les enjeux centraux que pose le principe de la *blockchain* qui ne se limitent pas à celui de l'anonymat, mais ouvrent sur les nouvelles formes de valeur, de gestion de l'identité, la confiance ou des formes de droit à venir dans des univers distribués.

Me voilà plongé pour les prochains mois dans des univers inédits, peut-être plus difficiles à pénétrer, mais dont l'exploration contribue petit à petit à esquisser de nouveaux horizons de recherche.

Études cartographiques

Chapitre 12. Archéologie des connaissances

« Il se peut bien – et encore, ce serait à examiner – qu’une science naisse d’une autre ; mais jamais une science ne peut naître de l’absence d’une autre, ni de l’échec, ni même de l’obstacle rencontré par une autre »,

Michel Foucault, *Les mots et les choses*, 1966.

Lorsque l’on me proposa de cartographier les acteurs et les projets de la Fondation pour la recherche sur la biodiversité (FRB), je sautai évidemment sur l’occasion, et pour trois raisons. La première paraît la plus évidente : nous allions pouvoir déployer dans un nouvel espace, *ailleurs* que sur le Web, nos instruments et notre méthode. Il s’agissait d’opérer la migration délicate de technologies et de compétences humaines vers l’univers cloisonné des bases de données documentaires internes. Pour le documentaliste de la fondation, cependant, notre arrivée ressemblait beaucoup plus à une invasion ! Mais chacun devina rapidement les plus-values que pouvaient représenter des cartographies des activités de la FRB et des communautés de recherche qu’elle avait pour mission d’animer. Pour nous, c’était le premier voyage hors du Web à destination, cette fois, d’un univers circonscrit et discret.

La seconde motivation de cette expédition tient en la qualité remarquable des informations contenues dans la base, toutes vérifiées et maintenues à jour lorsque le projet a débuté¹. La base constituait donc un véritable *trésor informationnel*, un capital qu’ignorent souvent à mon sens les organisations et qu’elles devraient valoriser. Pour la FRB, cette base est d’autant plus centrale que la fondation avait pour mission de mettre en relation et de fédérer les chercheurs et les structures impliquées dans un type de recherche par nature pluridisciplinaire et souvent international, la « biodiversité ». Cette base, en partie en accès public, a toujours fait l’objet de soins particuliers depuis sa création en 2005, en termes d’évolution technique comme de contrôle qualitatif des données. Elle contenait déjà en 2010 des informations sur plus de 4 300 chercheurs français et 1 200 étrangers qui publient dans des journaux internationaux à comité de lecture et/ou participent à des projets de recherche financés sur appels à projets compétitifs. Elle intégrait aussi à l’époque des renseignements sur 342 laboratoires de recherche français participant plus ou moins directement à la recherche sur la biodiversité, et 308 projets de recherche financés depuis 2005 sur ce thème.

1. Cédric Chavériat, Flora Pelegrin (responsable du pôle Stratégie et animation scientifiques), Fatima Fadil (chargée de mission portail et interfaces web) et Xavier Le Roux (directeur de la FRB).

Structure de données indexées dans la base de la FRB.

Base de données

Laboratoires	Projets	Acteurs
Les identifiants (nom*, abréviation, code d'unité)	Les identifiants (nom, acronyme)	Les identifiants (nom*, pré-nom*)
Les coordonnées	L'appel à projets	La fonction (chercheur, enseignant-chercheur, autre acteur)
Les contacts	Une date de début et une date de fin	Les coordonnées professionnelles
Une URL	Le pays du coordonnateur	Les contacts professionnels (téléphone, fax, email*)
Un résumé des activités de recherche	Une URL	Une URL
Les mots-clés	Un résumé	Un résumé des activités de recherche
Les acteurs qui en font partie	Les mots-clés	Les disciplines
La ou les institutions de tutelles	Les financements s'ils sont disponibles	Les mots-clés
Les structures qui la composent (les équipes de recherche...)	Les laboratoires participants	Les terrains d'études
La ou les structures dont elle fait partie (le département d'une université)	Les acteurs participants	Les écosystèmes étudiés
		Les texons étudiés

Avec autant de descripteurs et de croisements potentiels, la base de la fondation s'est avérée une véritable aubaine pour un cartographe ! On pouvait ainsi multiplier les angles de vue sur les formes et l'intensité d'un réseau de coopération sociale et professionnelle avec une précision que ne permettent pas les données web ou les publications scientifiques. En concentrant nos efforts sur des cartographies représentatives de la dynamique sociale ou collective engendrée par la construction de projets multidisciplinaires de recherche, nous avons éveillé l'intérêt que peut y porter une *gouvernance* (le terme est horrible !) dans son appétence pour les outils de gestion ou de mesure à long et moyen terme des effets de sa politique de mise en réseau et d'animation, la mission principale de la fondation. S'appuyer sur un état des lieux exhaustif des mises en réseau générées par une politique de labellisation des projets, voilà un chantier qui ne manquait pas d'intérêt pour un cartographe.

Mais la raison essentielle de cette expédition cartographique qui s'est déroulée sur plus d'une année tient en cette question : est-il possible d'esquisser la géographie intellectuelle générale sur laquelle tiennent la *recherche en biodiversité*, sa tectonique théorique, son architecture scientifique à titre de construction collective ? C'est le type de questionnement qui peut animer habituellement ceux qui connaissent ce champ de recherche de l'intérieur, à titre d'experts. Ou encore ceux dont le travail de recherche consiste à examiner les agencements théoriques qui président aux objets de

sciences, nouveaux ou anciens, comme en épistémologie et en histoire des sciences. En l'occurrence, c'est ce même désir qui nous animait donc, à ceci près, évidemment, que nous partions de données massives et capitalisées depuis plus de 5 années. Nous faisons alors le pari que l'analyse de données telle que nous la pratiquions pour les *web data* pouvait être appliquée aux bases qualifiées de connaissances. De notre point de vue, on pouvait considérer la base de la FRB comme un système d'indexation des traces de l'activité scientifique, qui s'amoncellent comme des couches temporelles dessinant des *patterns* remarquables et inaperçus jusque-là. Cependant, nous n'avions aucune garantie d'en trouver dans nos données et c'est en toute incertitude que nous avons commencé notre entreprise d'exploration, nourris par le désir de voir déconstruite, là devant nous l'architecture actuelle des connaissances qui se mettent en place autour d'objets complexes et nouveaux tels que la recherche sur la *biodiversité* comme d'autres pourraient le faire sur *l'environnement*, les *nanotechnologies* ou les *réseaux*. Tout dans cette entreprise me rappelait l'esprit de *L'archéologie du savoir*², l'œuvre magistrale de Michel Foucault.

La fondation nous avait demandé d'étudier la dynamique de coopération entre laboratoires français dans le cadre des réponses aux appels d'offres sur les cinq années précédentes. Cette demande ponctuelle reposait sur un travail de croisement des données-projets avec toute une famille d'attributs présents dans la base, comme les dates, les lieux, les objectifs de chacun des projets, les biotopes étudiés, les laboratoires ou les noms des chercheurs. Avec les données d'un côté et notre méthodologie venue du Web de l'autre, nous étions équipés pour répondre aux demandes de la FRB, mais aussi à des questions essentielles : comment s'organise donc ce vaste réseau d'acteurs coopérants, qui s'agrègent autour d'objets d'étude comme les *écosystèmes*, les *espèces animales*, la *génétique des végétaux* ? Selon quelle logique se croisent le chimiste, l'éthologue et le biologiste dans l'étude de biotopes qui s'étendent depuis les fonds marins, les milieux aquatiques continentaux, les zones cultivées jusqu'aux déserts, sous l'angle de l'histoire comme de la génétique ? Comment s'articulent autour de la biodiversité les approches et les disciplines (biologie, biochimie, agronomie, génétique, médecine, santé, écologie, histoire...) étant donné la complexité de l'objet ? Il se trouve que la BDD (Base De Données) de la fondation contenait le genre d'indices suffisant pour entamer un vaste travail de cartographie des acteurs et des thèmes de recherche sur la biodiversité, notamment parce que chaque chercheur peut définir et déclarer lui-même dans la base sa ou ses disciplines

2. M. Foucault, *L'archéologie du savoir*, Paris, Gallimard, 1969.

de rattachement. Ainsi, plutôt que de décrire les disciplines très générales auxquelles se rattachent les laboratoires, nous avons donc l'occasion d'analyser de près la dynamique d'agrégation de compétences personnelles dans les projets labellisés.

Les bases documentaires fermées n'incluent généralement pas de structures de graphes directement accessibles. Sur le Web, comme dans d'autres systèmes techniques (pensons aux aéroports et aux liaisons aériennes, ou encore au réseau électrique reliant des transformateurs), les liens sont directement dénombrables, donc représentables sous forme de graphe. Par exemple, quand on « *parse* » des pages web, on peut repérer automatiquement les liens hypertextes (*href*) et les lister dans un fichier de graphe. Dans une base fermée, les données documentaires ont besoin d'être transformées en structures de graphes. Si l'on réduit un système documentaire à un tableau classique de données, décrivant par exemple les projets d'une fondation, lignes et colonnes se croisent de multiples façons. On peut extraire de ce genre de structure informationnelle une série de graphes en exploitant un principe généralisé de redondance.

Image 12.2

Mots-clés	Projets	Acteurs
MC1, MC2, ...	P1	A1, A2, A3, ...
MC2, MC3, ...	P2	A1
MC1, MC4, ...	P3	A1, A3
MC3, MC4, ...	P4	A3, A4, A6, ...

Structure d'un jeu de données relationnelles.

La cartographie d'informations commence donc par une sorte d'art des croisements. Par exemple, un chercheur est impliqué dans un projet 1 (P1) ainsi que d'autres collègues, mais il est aussi *par ailleurs* impliqué (donc signalé dans le tableau de données) dans un projet 2 (P2), même situé à une autre date. De là, le principe de l'extension du procédé à la totalité des données pour produire une structure générale sur laquelle viennent opérer différentes formes d'analyse réseau, comme pour le Web. Ainsi, les projets de la fondation peuvent être classés ou analysés en fonction des chercheurs communs qui y participent comme en fonction d'autres dimensions (par les mots-clés décrivant l'objectif d'un projet, ce que l'on appelle à la FRB les *taxons*).

Image 12.3

Plusieurs réseaux peuvent être extraits d'un même jeu de données.

Notre croisement projets-chercheurs-compétences ou disciplines scientifiques ne représentait que l'une des nombreuses combinatoires possibles pour produire une structure de graphe. En effet, plus la table comporte de dimensions ou de colonnes, plus les occasions de croiser les données – donc de les cartographier – sont nombreuses. On approche ainsi de la notion d'*information complexe*, autrement dit *multi-dimensionnelle* et interrogeable sous de nombreuses facettes. Les objets complexes constituent encore aujourd'hui un fameux terrain d'investigation scientifique et, avouons-le, notre méthode pour générer des structures de graphes n'est pas des plus avancées au regard des procédures en mathématiques appliquées, en économétrie ou en sémantique. Plutôt que de générer des espaces en n dimensions pour saisir, analyser et comparer les multiples dimensions d'un jeu de données, la cartographie à base de graphes ne peut reposer, au mieux, que sur le croisement de deux, voire trois dimensions au maximum et encore, successivement. Nous ne mobilisons donc pas ces sortes de « boîtes noires » hautement « abstraites » : la rentabilité ou la plus-value scientifique de la cartographie vient *en aval* quand on projette sur une structure de graphes les métriques réseau qui nous font apercevoir les informations sous forme de topologies. Dans le travail pour la FRB, la réduction de la complexité de l'information s'est effectuée dans le croisement systématique, les dimensions prises deux à deux pour trouver des angles pertinents et apercevoir des *patterns* remarquables (lorsque je saisis le résultat du croisement sous la forme d'un *cliché*). Évidemment, pour notre approche, la rubrique « projets labellisés » semblait constituer le ciment qui fait tenir ensemble aussi bien les chercheurs, leurs compétences, leurs laboratoires ou leurs biotopes favoris que les concepts, les disciplines, les mots-clés ou les descripteurs de contenus. Par définition collaboratifs et multidisciplinaires, ils *écrivent* sous forme de traces matérielles (des documents) la dynamique sociale d'un réseau de recherche comme la richesse et la diversité des savoirs qu'ils mobilisent.

Image 12.4

Laboratoires et structures par régions

Les 343 laboratoires impliqués dans les 258 projets labellisés par la FRB ont ici été regroupés par régions administratives françaises (hors DOM-TOM regroupés en un seul ensemble). A gauche, la vue circulaire où les régions ont été distribuées en fonction du nombre de laboratoires impliqués par ordre décroissant. A droite, distribution géospatialisée. La taille des liens, dans les deux cas, sont fonction du nombre ou de l'intensité de collaborations entre laboratoires des différentes régions.

Réseau de collaborations entre laboratoires identifiés par régions

Les projets de recherche sur la biodiversité française dans le cadre d'Appels à propositions de recherche permettent d'établir une carte représentant le réseau de collaborations entre les différents laboratoires de recherche.

Les AAP suivantes ont été pris en compte :

ANR-Biodiv-2009 (12)	ANR-Biodiv-2006 (15)	ANR-Biodiv-2008 (15)	ANR-Biodiv-2007 (16)
ANR-SYSTERIA-2008 (15)	ANR-SYSTERIA-2009 (15)	ANR-SYSTERIA-2010 (11)	BRQ-PO-2009 (20)
BRQ-PO-2007 (12)	FRS-CESU-2010 (3)	FRS-Genovo-2009 (13)	FRS-Genovo-2010 (11)
FRS-Genovo-2007 (17)	FRS-Genovo-2008 (17)	MAUR-Genovo-2009-2007 (7)	MEBIO-Genovo-2009 (15)
MEBIO-Genovo-2010 (9)	MEBIO-LTERA-2007 (1)	MEBIO-LTERA-2009 (2)	NPCFRS-Biodiv-2010 (12)

- Ile-de-France
- Languedoc-Roussillon
- Bretagne
- Provence-Alpes-Cote d'Azur
- Midi-Pyrenees
- Rhone-Alpes
- Aquitaine
- Autres régions (incluant les DOM-TOM)

L'étude des collaborations entre les laboratoires français autour des problématiques de biodiversité a constitué une part importante du travail de cartographie. Les laboratoires de recherche ont été regroupés par région (vues circulaire et géoréférencée à gauche) pour étudier de façon macroscopique les réseaux de collaboration. Dans des vues plus précises où figure chacun des laboratoires, les nœuds ont été anonymisés pour éviter les polémiques potentielles (un nœud est d'autant plus volumineux qu'il participe à de nombreux projets).

La première série de cartographies réalisée pour la fondation constitue une sorte de catalogue, presque un atlas, dont les vues participent à une forme de *dashboard social* qui correspondait aux demandes de la fondation. Ces cartographies constituent des éclairages successifs et complémentaires sur la base, mobilisables par la suite à titre d'instruments d'évaluation des activités de la fondation (nos cartes couvraient sept appels à projets et 286 projets labellisés) ou d'instruments d'aide à la décision (pour identifier les pistes émergentes de recherche). Nous n'étions plus sur le Web, libres de nos directions expérimentales, mais contraints par les cadres d'une organisation, de ses services et de ses objectifs. Les données de la base, une fois transformées en cartes ou en indicateurs, pouvaient alors être mobilisées au service de la stratégie d'animation d'un immense réseau d'acteurs (plus de 4 000 chercheurs identifiés), notamment dans des périodes de recherche d'experts ou d'activation de réseaux de coopération à l'échelle européenne dans le cadre des appels à projets. Pour la première fois, les *patterns* relationnels qu'exhibaient nos graphes allaient être mobilisés au service d'une analyse stratégique, ce moment où la structure de l'information n'est plus un objectif de recherche, mais un vecteur ou un instrument accompagnant la décision ou l'évaluation.

Image 12.5

L'identification et la description des communautés de chercheurs ont constitué une part importante du travail de cartographie. Ici, les chercheurs sont reliés entre eux en fonction de leur(s) participation(s) à différents projets de recherche. Ces derniers contiennent une description du type de biotope qu'ils analysent. En croisant les données-chercheurs avec les descripteurs de biotope, on peut produire une carte des biotopes et des communautés de chercheurs qui travaillent dessus.

Pendant que la production des vues cartographiques se poursuivait, notre projet de cartographie générale des savoirs scientifiques mobilisés autour de la question de la biodiversité commença réellement. Les affiliations de disciplines déclarées par les chercheurs eux-mêmes (en rappelant que la BDD de la FRB peut être utilisée comme un instrument de recherche de partenaires potentiels) ont constitué le principal élément de notre entreprise cartographique. Si l'on possède les noms des chercheurs impliqués dans la totalité des 286 projets labellisés, il devient alors possible d'agréger autour de chacun d'eux les différents champs scientifiques et de les traiter sous forme de graphe pour chercher à en comprendre l'articulation ou la complémentarité à grande échelle. Les données de la FRB associées à notre méthode de travail ouvraient donc directement sur la dynamique sociale des projets de recherche collaboratifs et pluridisciplinaires, comme une sorte de logistique théorique associée à chacune des problématiques traitées dans les projets labellisés. Nous étions arrivés à un niveau de granularité descriptive beaucoup plus fin que ne le sont les « grandes » catégories des éditeurs comme le Web of Science ou Scopus. Et surtout, nous avions ainsi accès à un ensemble d'indicateurs forts sur les complémentarités de disciplines scientifiques telles que les associent les chercheurs dans leurs activités.

Notre objectif, ici, n'a pas été de produire une carte générale des sciences, ni même d'y situer la recherche en biodiversité (ce que l'on peut faire à partir de grandes bases comme le Web of Science), mais seulement de remonter « en abstraction » depuis le niveau granulaire des acteurs, en passant par les mécanismes de leurs interactions, vers la façon dont s'agencent les grands continents de la connaissance scientifique et de sa distribution en disciplines à travers notre analyse des projets labellisés par la fondation. Il s'agissait donc de déformer l'arborescence générale des disciplines et de ses catégories en introduisant des masses de liens transversaux pour espérer obtenir une structure de graphe interprétable.

Disciplines de la recherche en biodiversité

Physique	Médecine, santé	Biologie	Biologie des populations et écologie	Sciences agronomiques	Sciences de la société
Physique et science des matériaux	Épidémiologie	Biologie cellulaire	Biologie de l'environnement	Agronomie	Science de gestion
	Pharmacologie	Biologie moléculaire	Écologie	Phytotechnie	Science économique
	Thérapeutique	Botanique	Écologie urbaine	Agriculture	Droit public
Chimie et science des matériaux	Maladies infectieuses	Cladistique	Écologie du paysage	Sciences vétérinaires	Sciences politiques
Chimie		Génomique	Biologie marine	Productions animales	Anthropologie
Chimie organique, minérale, industrielle	Sciences et technologies de l'information et de la communication	Biomolécules	Hydrobiologie	Productions végétales	Ethnologie
Chimie théorique, physique, analytique	Traitement du signal et des images	Biologie des organismes	Écophysiologie	Sciences et techniques des pêches	Sociologie
	Informatique et ses applications	Biologie des populations	Génétique quantitative	Sciences forestières	
Mathématiques et leurs interactions		Biologie de la reproduction	Génétique des populations	Sciences de la terre et de l'univers, espace	Sciences humaines et humanités
Mathématiques	Sciences pour l'ingénieur	Biologie du développement	Biogéographie	Sciences de la terre et de l'univers	Langues et littérature
	Acoustique	Endocrinologie et métabolisme	Phytogéographie	Climatologie	Éthique et déontologie
	Bio-mécanique	Entomologie	Biologie appliquée	Géologie	Épistémologie,
	Bio-ingénierie	Evo-devo	Biologie animale	Géophysique	histoire des sciences et des techniques
	Génie civil	Ichtyologie	Biologie végétale	Géochimie	Aménagement de l'espace et urbanisme
	Mécanique des matériaux, des structures et des surfaces	Malacologie	Biotechnologies agroalimentaires	Hydrologie	Géographie physique, humaine, économique et régionale
		Mycologie	Biologie de la conservation	Paléontologie	histoire et civilisations
		Phylogénétique	Écotoxicologie	Pédologie	Préhistoire
		Physiologie	Bioprocédés et microbiologie	Physique et chimie de l'océan	Sciences de l'information et de la communication
		Toxicologie	appliquée		
		Mammalogie	Bioclimatologie, sciences de l'environnement		
		Microbiologie			
		Ornithologie			
		Parasitologie			
		Systématique			
		Taxonomie			
		Virologie			
		Zoologie			
		Ethologie			

Liste de disciplines utilisée comme base pour construire le réseau.

La phase initiale de notre projet consistait à supposer l'existence d'une *forme* (interprétable, compréhensible, presque rationnelle) d'organisation des savoirs, tout en étant habités par quelques doutes sur la consistance du phénomène : verra-t-on apparaître un univers morcelé de disciplines scientifiques ou, au contraire, un seul ensemble très densément connecté ? Contiendra-t-il des îlots ou des continents, éventuellement organisés selon des directions ? Les données de la dynamique sociale d'agrégation des chercheurs sur plusieurs années permirent de redistribuer les 96 disciplines selon la logique d'association de compétences des 1 700 chercheurs sur 296 projets qui révéla presque exactement le type de graphe que nous attendions :

Image 12.7

Réseau des disciplines (spatialisation avec Force Atlas 2).

Comme certains se le sont écrié en regardant la spatialisation obtenue avec Gephi, « un plat de spaghettis ! ». Au-delà de la blague, le plat de spaghettis augure, pour l'habitué que je suis, d'un espace relativement organisé, où tous les éléments sont *connexes*, appelés peut-être à s'agencer d'une certaine façon ou comme des indices des multiples *complémentarités* entre savoirs scientifiques différents et mobilisés autour d'une question. Certes, nous sommes là en face de quelque chose qui n'a plus rien à voir, semble-t-il, avec la *liste*, la *formule* et le *tableau* dont parlait Jack Goody³. À certains égards, nous n'en sommes pas très éloignés non plus avec une telle spatialisation : les nœuds ont des grosseurs variables selon le nombre de fois où la discipline est associée à une autre et l'on peut donc en faire la liste ordonnée (principe du

3. Jack Goody, *La raison graphique. La domestication de la pensée sauvage*, Paris, Minuit, 1979.

ranking). Les couleurs renvoient aux 13 sous-ensembles dans lesquels elles étaient regroupées dans l'arborescence de départ (les couleurs dites « chaudes » pour les sciences du vivant dans la partie gauche). La structure visuelle sous-jacente, en revanche, peut en laisser certains surpris par l'apparition transversale d'autant de liens ou de connexions, dans un univers que ne permettait plus d'apercevoir l'arborescence des disciplines de départ. L'espace graphique et logique y est d'un certain niveau de complexité, comme si l'on passait à un niveau supérieur d'abstraction. L'arbre hiérarchique des disciplines semblait ainsi s'être dilaté dans un espace de propriétés statistiques où les notions de *distances*, de *diamètre*, de *densité*, de *chemins*, de *centre* et de *périphérie* vont alimenter désormais l'interprétation.

Indéniablement, un nombre réduit de disciplines issues des sciences du vivant forment un cœur dense de relations, donc très souvent associé aux projets de la fondation. Au cœur de la question de la biodiversité figure donc la question du vivant ou de la « vie » au sens biologique. Autour de la *biologie*, un rôle majeur semble joué par une série de disciplines très reliées entre elles, mais aussi reliées, depuis le cœur, à toutes les autres (*biologie des populations*, *biogéographie*, *génomique*, *écologie*, *biologie de l'environnement*). À ce niveau, la *biologie* semble s'étirer en différentes branches dont les spécialisations ou les directions ne nous apparaissent pas encore. La partie droite du cœur contient au moins deux types de disciplines : *agriculture*, *sciences forestières*, *agronomie* et *écologie du paysage* d'un côté ; de l'autre, *l'économie*, *l'ethnologie*, la *sociologie* et la *géographie physique, humaine, économique et régionale*. Ce rôle majeur peut-être vérifié via une visualisation circulaire : il s'agit du même jeu de données, mais dont les éléments sont projetés, cette fois, sur un cercle et regroupés par catégorie. Le principe permet d'étudier la distribution générale de la connectivité entre éléments, notamment les zones de grande densité de liens. On y aperçoit nettement la densité importante qui associe *biologie de l'environnement* et *biologie des populations* via un lien rouge dont l'épaisseur rappelle que, de très nombreuses fois dans les projets soutenus par la fondation, les deux champs scientifiques sont intensément associés. On y aperçoit aussi les liens plus secondaires, mais importants qui relient science économique et géographie physique et humaine au reste des sciences du vivant.

Image 12.8

Réseau des disciplines (spatialisation circulaire).

Cet indicateur circulaire de distribution et de la densité de la connectivité entre disciplines n'épuise pourtant pas les pistes que nous propose la première vue générale dans Force Atlas 2. Les forces d'attraction et de répulsion de l'algorithme ont généré cette vue générale qui a des vertus particulières. À l'examen, elle fait apparaître au premier abord quelques éléments curieux pour un néophyte comme la « cladistique » (ou l'étude de la classification des êtres vivants, selon leurs relations de parenté, dans un cadre évolutionniste) ou encore « EvoDevo » (génétique évolutive du développement) sans oublier la « Mammalogie » (étudier les origines, le comportement, le régime alimentaire, la diversité génétique et la dynamique des populations des mammifères) que je ne connaissais pas. Mais il y a plus : elle situe (ou « contextualise ») chacune des disciplines les unes par rapport aux autres, notamment à partir des chemins, plus ou moins denses, dessinés par les liens. Quand on en suit quelques-uns depuis le cœur, ils s'achèvent dans l'une des branches périphériques où sont positionnées essentiellement les disciplines moins mobilisées. Elles s'y regroupent parce qu'elles partagent en partie le même type de distribution de la connectivité (elles sont liées ensemble localement, mais, de surcroît, elles sont reliées aux mêmes types

de disciplines dans le cœur). Quand on est néophyte dans un domaine de recherche, il est difficile d'apercevoir tous les accidents historiques (par exemple, deux disciplines qui se rapprochent autour d'un nouvel objet), sociaux (des chercheurs et des laboratoires qui aiment à travailler ensemble, ou pas) ou tout simplement l'arrière-plan théorique (tels hypothèses ou modèles théoriques viennent se loger à tel endroit de la carte) qui aboutit à ce type complexe d'agencement des 96 disciplines. Des experts se sont prononcés de nombreuses fois sur cette carte lorsque nous l'avons imprimée en A0, un format idéal de discussion collective, et il faudrait qu'une sociologie des sciences vienne ici relever tous les événements qui ont contribué à livrer une telle géographie, depuis les témoignages des chercheurs jusqu'à l'analyse des documents présents dans la base ou à la fondation. Pour un explorateur de *data*, voire un cartographe de l'information comme moi, il fallait réduire les propriétés essentielles de cette carte en principes intelligibles et les synthétiser sous une forme concentrée. Une boussole des disciplines, voilà ce qu'il fallait.

On peut faire l'hypothèse que ce type d'agrégation des disciplines par le jeu d'attraction/répulsion permis par l'analyse des liens (et donc des compétences humaines qui y concourent) ne devrait avoir rien de fortuit et tout d'un programme historique, comme une photographie ou une empreinte de ce que peut être un champ de recherches interdisciplinaires. Cela ne veut pas dire que les 96 disciplines identifiées sont à considérer de façon équivalente, ne serait-ce que par la hiérarchie introduite par l'analyse de leur occurrence (une discipline apparaît plus souvent dans les projets qu'une autre) ou de leur co-occurrence (une discipline est plus ou moins souvent liée dans les projets à une autre). À ce titre, une dizaine de disciplines constitue un cœur remarquablement concentré de relations (dont l'écologie, la biogéographie et la biologie des populations), comme si la question de la nature de l'homme, depuis la génétique jusqu'à la géographie, imprimait à ce vaste ensemble qu'est la biodiversité, un centre de gravitation très fort. Ce n'est pas la seule dimension spatiale ou physique de la vue générale : les « branches » semblent indiquer, dès le parcours attentif des premiers chemins, une « logique » ou une « raison » qu'une forme d'interprétation doit pouvoir expliquer ou justifier. Avec ce type de projection, les distances (ou les éloignements) visuelles ou statistiques importent, mais aussi l'orientation des éléments. Cette sorte de géologie épistémologique que nous cherchions à isoler réclamait des attracteurs périphériques, presque des frontières, qui nous auraient indiqué les limites de la recherche sur la biodiversité. Notre boussole des disciplines (ou boussole thématique) devait permettre de naviguer dans l'espace de la recherche sur la biodiversité et d'en apercevoir les grands traits d'un seul coup d'œil ou d'y situer un champ de recherche précis. À bien y réfléchir, elle pourrait aussi permettre de situer des chercheurs, des laboratoires ou des projets (par exemple les futurs projets abellisés),

par superposition ou comme si on la recouvrait d'un calque. Ou bien encore, remise à jour chaque année, on aurait pu lire les évolutions possibles ou probables en fonction du temps.

Plusieurs réductions ont ainsi été tentées pour identifier les orientations majeures du système des disciplines et poser les grandes lignes organisatrices de notre boussole. L'une, en particulier, a consisté à réduire comme en cartographie (géographique) le territoire de départ en carte sur le principe du positionnement des 13 grandes catégories de disciplines. Le plan qui se dessine alors part des sciences pour l'ingénieur vers la médecine *verticalement* et de la chimie vers les sciences humaines et vers les mathématiques *horizontalement* :

Image 12.9

Réduction en 13 grandes catégories.

Une autre réduction a été opérée en utilisant cette fois le principe du zonage. Cette version synthétique de la vue principale indique clairement une partition gauche-droite, de la biologie et la biochimie à gauche aux sciences humaines à droite et avec une colonne vertébrale qui va de l'ingénierie en haut aux sciences de la terre et des espaces en bas.

Image 12.10

Visualisation des catégories par zonage.

Les deux procédés ne permettent pourtant pas encore de saisir la logique d'association des disciplines. Si l'on peut déjà admettre la bipartition manifeste de gauche à droite, entre la « biologie » (j'aurais tendance à dire « sciences du vivant et de la nature ») et les SHS (au sens large : l'homme et l'univers tels qu'ils sont produits culturellement, historiquement et géographiquement), il faut pouvoir justifier, en détail, de l'enchaînement des disciplines.

Image 12.11

Thérapeutique < Chimie organique, minérale, industrie < Toxicologie < Biologie des organismes < Biogéographie < **Ecologie** > Agronomie > Phytotechnie > Informatique et applications > Sciences de l'information et la communication > Langues et littératures

Organisation des disciplines de façon linéaire.

L'étude de chemins particuliers peut y aider. Si l'on en revient à la carte générale (donc à une structure de graphe qui est fondamentalement affaire de liens) et que l'on s'y promène en partant des deux disciplines les plus éloignées l'une de l'autre sur un plan horizontal (la « Thérapeutique » à gauche, « Langues et Littérature » à droite), on peut se rendre compte que le graphe contient comme une déclinaison des disciplines qui, de proche en proche et telles qu'elles sont liées dans des projets effectifs menés par les chercheurs, nous fait passer (et sans hiérarchie puisque les liens sont non dirigés) de l'univers des « corps » biologiques (substrat naturel de la biodiversité) à l'univers abstrait et contingent de la culture et de ses productions (ici, la littérature, mais en passant aussi par d'autres productions comme l'agriculture, les sciences de l'ingénieur ou les biotechnologies que l'on peut regrouper sous le label de « services écosystémiques »). Dans ce type de parcours, de gauche à droite et inversement, l'écologie et la biologie des populations constituent, non plus un cœur dense, mais des *disciplines-pivot*.

Or, une autre lecture est possible, ordinale cette fois et très complémentaire. Sur le schéma, les deux ensembles majeurs que sont la « Biologie » et les « Sciences de la société » sont reliés via une ligne de fracture dont « l'Écologie » et la « Biologie des populations » sont encore le pivot : vers le haut, se profilent un ensemble de disciplines qui traitent de l'« environnement » au sens large, tel qu'il est produit, cultivé, exploité (sciences et techniques) ou tel qu'il impose aussi ses contraintes (bioclimatologie). Ainsi, la distribution des disciplines, dans la moitié supérieure de la carte, trouve sa logique : de part et d'autre d'une ligne verticale, Écologie/Botanique/Climatologie, on trouvera (sans surprise) d'un côté les disciplines qui étudient « l'environnement » sous l'angle de la culture (Écologie urbaine, Aménagement de l'espace et urbanisme, Histoire, Sciences politiques...) et, de l'autre côté, les disciplines qui étudient l'environnement sous l'angle de ses rapports au vivant (Écotoxicologie, Biotechnologie végétale, Endocrinologie et Métabolisme...). Dans la partie inférieure de la carte, la distribution obéit aux mêmes principes, mais au regard des

espèces vivantes qu'il faut décrire, inventorier, analyser en tant que telles, or aussi dans leurs relations réciproques à titre de systèmes (ou « d'écosystèmes »). En bas, à gauche, un ensemble gouverné par la biologie des interactions (virologie, parasitologie, biotechnologies animales...) et, à droite, l'ensemble complémentaire des disciplines qui étudient les espèces vivantes (animales et végétales) sous l'angle, cette fois, de leur production historique (Sciences forestières, Agriculture, Sciences vétérinaires, Bio-ingénierie...).

De là, on peut imaginer dans le paysage intellectuel qui se dessine la recherche sur la biodiversité comme une intersection dont les axes orientés fixent l'espace comme les limites.

Image 12.12

Organisation des disciplines en boussole.

La boussole finale, le « plan » si l'on veut de l'architecture disciplinaire de la recherche en biodiversité contient des axes, déployés depuis le centre vers la périphérie. Les savoirs scientifiques s'enchaînent pour couvrir tous les champs possibles de questionnement (du moins, dans la version actuelle de la configuration) où la question de la vie (et en son centre l'homme) se décline en fonction des autres espèces et des interactions (bas) et d'un environnement global où le végétal est inclus (haut), mais aussi, simultanément, en fonction des propriétés biochimiques d'un être naturel jusqu'au conditionnement culturel propre à l'homme. Le titre, l'écosphère, vient confirmer l'effet de clôture de cette boussole.

Un tel effet de totalité pourrait être mobilisé efficacement pour montrer à un jeune public toutes les implications du concept de « biodiversité », en particulier la façon dont l'homme s'insère dans un univers vivant qu'il contribue, par ailleurs, à produire

lui-même. Au-delà de la valeur pédagogique de l'effet de circularité que renforce la forme englobante de la boussole, il reste que l'on ne sait pas à quoi en attribuer les propriétés : aux *patterns* sociaux qui semblent receler les *data* ou bien aux artefacts techniques et méthodologiques mobilisés ? D'un côté, ce résultat final issu d'une suite de transformations (« d'un support à l'autre », dit Bruno Latour⁴), depuis la table de données de départ en passant par les modifications successives dont le graphe et sa spatialisation n'ont été que le pivot central. De l'autre, l'enchaînement des disciplines et leur agencement général représentent un *pattern* organisationnel trop fort pour n'être pas considéré comme *constituant*.

En épistémologie, en histoire ou en sociologie des sciences, on retrouve là la configuration classique d'un débat infini, que l'effervescence actuelle autour des *big data* a ranimé de façon intense. Ces formes statistiques traduites en boussole sont-elles un mirage né de nos instruments ? Sinon, d'où viennent-elles ? D'un savoir *non su*, d'une organisation implicite à laquelle chacune des activités scientifiques se plie sans pouvoir l'énoncer ? On pourrait aussi envisager la question de façon plus prosaïque, en raisonnant en termes de pertes et de gains. La mutation des données d'origine sous forme de listes brutes jusqu'à la boussole finale passe par une succession de phases où l'on a construit une table multidimensionnelle dont on a extrait seulement deux dimensions, on les a ensuite croisées pour produire une structure de graphe, visualisée avec un algorithme tout à fait particulier (le Force Atlas 2 développé dans le contexte des données web), pour finalement examiner quelques principes structurants essentiels à une vue organisée, justifiable éventuellement à travers un discours explicatif *ad hoc*. Les pertes semblent donc incommensurables depuis le *data set* de départ, sans même prendre en compte le contexte humain et technologique particulier qui était le mien au moment de l'étude (c'est-à-dire tout à fait relatif). Au contraire, envisager les gains d'un tel processus de transformation oblige à considérer la boussole comme le produit d'un enrichissement continu de l'information de départ qui permet de cerner de façon synthétique les grands axes d'une organisation spécifique des savoirs scientifiques, structurés autour d'un domaine nouveau et multidisciplinaire. Sur ce point, les gains sont énormes, aussi bien en termes cognitifs (une carte des disciplines qui assure l'aide à la navigation dans des masses d'informations qui s'accumulent année après année) que sociaux ou politiques pour accompagner la fondation dans l'animation de ses nombreux réseaux de coopération.

C'est là, à mon sens, la véritable plus-value générée par notre démarche : le passage des masses de données aux *smart data*, en particulier la traduction en indicateurs graphiques de propriétés statistiques massives pour déboucher sur des instruments de synthèse tournés vers la construction de l'interprétation, voire de la décision.

4. B. Latour, *La clef de Berlin et autres leçons d'un amateur de sciences*, Paris, La Découverte, 1993.

En accumulant les occasions de produire ce genre de boussoles, en stabilisant une méthode, on pourrait même penser à automatiser la procédure, masquant ainsi sous forme d'une « boîte noire » toutes les opérations artisanales de transformation des *data* dont chacune arrêtaient auparavant le regard critique du sociologue ou de l'épistémologue. Ce projet de cartographie mené de façon artisanale tout au long de l'année 2010, nourri d'essais successifs et de modifications régulières de nos petites chaînes de traitement de l'information, contenait déjà en germes toute une série d'automatisations potentielles représentatives de l'ère actuelle gouvernée par la quête généralisée des *patterns* statistiques dans le *big data*. C'est à la même époque que je voyais se développer autour de moi des dispositifs dédiés à des masses analysées en temps réel et restituées sous forme de cartographies (le service InMaps de LinkedIn sur lequel travaillait Mathieu Bastian, les projets d'interfaces développées par Linkfluence qui allaient devenir Radarly, les avancées méthodologiques et techniques de Mathieu Jacomy sur la cartographie des controverses...). À petite échelle, et lentement, cette transformation progressive des *data* en boussole synthétique illustrait déjà une forme de quête généralisée des *patterns* contenus dans les bases de connaissances, mais aussi sur le Web. Nous ne faisons là que participer à notre modeste niveau à la formidable accélération des procédures de traitement de l'information au cœur du Web des données et des services associés qui émergeait de façon généralisée autour des années 2010 et dont les *data sciences* ont depuis fait leur objectif principal. Ce mouvement autour des *data* semble largement associé aux masses et à nos capacités décuplées de calcul sur les machines, mais sa figure principale m'apparaît résider dans la forme elle-même des *data processing* : ce que j'avais été amené à définir comme des chaînes de traitement de l'information, linéaires et modulaires, se transformaient à grande échelle sur le Web en boucles continues d'enrichissement où les *data*, à peine extraites du réseau, puis croisées, analysées, comparées, y retournaient enrichies en flux. Ce second âge des *data*, où les boucles de transformation continue remplacent petit à petit les processus linéaires de traitement, suppose la rénovation significative de nos outils d'analyse et de contrôle, un défi de plus pour la cartographie d'information à venir.

Chapitre 13. Révélations

« Nous avons dessiné alors un vaste schéma, que j'ai toujours avec moi. Il fait huit mètres une fois déplié. Il serpente depuis le bureau d'un directeur des hydrocarbures d'Elf, jusqu'à l'ouverture de comptes obscurs alimentés par le Gabon, aux mains d'Omar Bongo : quarante ans de pouvoir et une difficulté récurrente à distinguer sa tirelire et sa famille d'une part, le budget de l'État et le gouvernement d'autre part. J'emporte souvent ce schéma avec moi, au fil des rendez-vous. Je l'étale sur les tables, un peu comme un capitaine au combat sort ses vieilles cartes. Les positions ont sans doute varié, les techniques de camouflage se sont sophistiquées, mais le système est là : les tyrans sont des amis, que la France a placés au pouvoir et dont elle protège la fortune et l'influence par de vastes réseaux de corruption ; en échange ils veillent sur les intérêts et les ressources des entreprises françaises venues creuser le sol. Tout ce beau monde a intérêt à ce que rien, jamais, ne stimule ni les institutions ni l'économie des pays », Eva Joly, *La force qui nous manque*, 2007.

« ARGENT PUBLIC : Voici comment des fonctionnaires et des ex-fonctionnaires profitent des mandats de la Confédération. Un petit Atlas du copinage ». En ce matin du 25 novembre 2012 en Suisse, les termes utilisés par les éditions du *Matin Dimanche* et du *SonntagsZeitung* étaient forts et accusateurs¹. Je me précipitai sur les articles dès que l'on me signala leur diffusion sur le Web : l'enquête des journalistes était accompagnée d'une carte du réseau de « contacts amicaux » entre des acteurs qui semblaient « organiser » la publication comme l'attribution de marchés publics de la Fédération helvétique à leur profit². L'article m'est tout de suite apparu comme un exemple remarquable de *data journalism*. Signé de J. Schmidli, M. Stoll et T. Plattner, il étonne par la précision dans la description des dysfonctionnements du système d'attribution des commandes de la Confédération helvétique et la cartographie réalisée par Martin Grandjean constituait un exemple techniquement abouti et enthousiasmant de visualisation de données. Comme le rappelle l'article, le *data journalism* met « à disposition de la Confédération » cet outil qu'est la cartographie d'information « pour lutter plus activement contre la corruption et le copinage ! ». Si les mots décrivent, le graphe réalisé avec Gephi (évidemment) frappe parce qu'il atteste de la réalité d'un phénomène qui semble massif et systématique : le marché public helvétique est « noyauté par la très grande mobilité de ses hauts fonctionnaires, qui, au passage du public au privé (ou l'inverse), font profiter leurs entreprises de la manne gouvernementale ».

1. <https://humanitesdigitales.files.wordpress.com/2012/11/lmd.pdf>.

2. <http://web.archive.org/web/20170315085705/http://owni.fr/2012/11/29/la-nebuleuse-suisse-des-copinages-public-privé/>.

Image 13.1

Le réseau de Martin Grandjean dans « ARGENT PUBLIC : Voici comment des fonctionnaires et des ex-fonctionnaires profitent des mandats de la Confédération. Un petit Atlas du copinage ».

Les articles des journaux suisses illustraient remarquablement la façon dont la transformation d'une information sous forme de réseau venait conquérir des niches depuis son terrain d'origine, le Web. Au moment où paraissent la cartographie et les deux articles associés, bon nombre de compétences humaines, de méthodes et de technologies ont déjà migré depuis l'exploration du Web vers l'étude de silos locaux d'informations, et particulièrement des silos d'où il serait possible d'extraire des graphes concernant des acteurs ou des organisations, ce que l'on appelle des graphes d'*entités nommées*. Sur le Web, de grands acteurs industriels comme LinkedIn et Facebook en construisent aussi, et de très vastes puisque ce sont par nature des réseaux sociaux. Mais il est possible, en modifiant les technologies web et les ingénieurs qui les conduisent, d'aller aborder des bases locales d'informations ou de connaissances, capitalisées de longue date et richement contrôlées. Parfois, cette migration n'est qu'une revisitation comme lorsque l'on produit des graphes d'inventeurs ou de technologies à partir des bases de brevets ou des champs disciplinaires, des revues ou des auteurs dans des bases bibliographiques comme Scopus ou le WoS

– le Web of Science de Thomson Reuters. Dans ces deux domaines, de nombreux travaux à base de graphes existent depuis les années 1960, notamment aux États-Unis³. C'est pourquoi l'arrivée de méthodes et de technologies nouvelles imaginées pour le Web est souvent vécue par les spécialistes de ces domaines comme une forme d'intrusion ; la diversité et la robustesse des nouveaux algorithmes (HITS, PageRank, Modularity, OpenOrd notamment) comme l'étendue des masses traitées permettent de « voir » les documents juridiques et les publications scientifiques sous de nombreux aspects inédits⁴. Mais il existe une foule d'autres terrains informationnels à explorer, entièrement nouveaux. Et parfois, cela a tout d'une prise d'assaut si l'on pense à certains types de données libérées dans le mouvement actuel de l'*open data*. C'est là que les instruments cartographiques développés pour l'exploration du Web et de sa structure hyperliée font merveille, surtout quand il s'agit d'extraire des données rattachées à des « entités nommées », aux interactions qui les lient en réseaux parfois inextricables. Autrement dit, des *patterns* relationnels qui peuvent être interprétés comme des formes implicites d'organisation sociale, dont certaines semblaient *cachées* jusqu'à présent. Si l'on associe cette sorte d'appétence naturelle que nous partageons tous pour les révélations (en particulier si elles concernent des événements ou des personnages de premier plan), aux capacités actuelles de traitement de l'information (sur le Web ou ailleurs), on comprend rapidement les pouvoirs de l'approche par graphe. Tout ce qui participe de la dynamique sociale à titre d'activité organisée (entreprises, marques, ministères, établissements publics, partis politiques, associations, mouvements, familles et proches...) peut un jour se trouver cartographié (ce qui n'est déjà pas anodin sous certains aspects), et sous un jour nouveau à partir d'informations sensibles et/ou de divers silos informationnels croisés entre eux. Et publiques la plupart du temps. Le procédé n'est pas nouveau au regard de l'histoire des médias ou du journalisme d'enquête, mais la cartographie permet d'embrasser l'ampleur d'un phénomène, objectivant d'un coup un ensemble des mécanismes collectifs complexes... et parfois très discrets jusque-là.

Lors de la parution des deux articles suisses, cela fait déjà quelque temps que certains des ingénieurs qui m'avaient accompagné jusqu'alors dans l'exploration du Web se sont mis à occuper certaines niches de services autour de l'information stratégique, des outils de *monitoring* ou de l'aide à la décision. Certaines affaires qui ont défrayé la chronique au cours des années passées ont été pour eux l'occasion de communiquer sur leurs compétences techniques en accompagnant quelques fameux « lanceurs d'alerte ». Les premiers auront été les fondateurs de la société Linkfluence

3. Voir le chapitre 4 des *Carnets cartographiques*, « Cartographier la science », <https://ateliercartographie.wordpress.com/2016/12/22/cartographier-la-science-le-chapitre-4-des-carnets-cartographiques/>.

4. Voir par exemple : <http://web.archive.org/web/20180308055220/http://lexmex.fr/> développé par Jacques Verrier en 2012 et la vidéo associée sur l'évolution du Code civil.

(fondée en 2005, spécialisée aujourd'hui dans le *monitoring* du web social et dans la mesure de l'influence des « entités sociales »). Ils ont ainsi équipé des équipes de journalistes dans l'analyse et la publication sous forme de cartographies des câbles diplomatiques américains dans l'affaire WikiLeaks. Le dispositif CableGate Radar⁵ hébergé sur le site du journal *Le Monde*, très abouti techniquement, permet ainsi de naviguer publiquement dans les dépêches secrètes des ambassades américaines dans le monde en faisant des recherches croisées par mots-clés, zones géographiques, segments temporels ou personnes citées :

Image 13.2

Le CableGate Radar de Linkfluence.

D'autres sont partis travailler à l'exploration des données sociales chez LinkedIn. Plus récemment, l'Atelier Iceberg⁶ s'est spécialisé dans la visualisation des données internes des entreprises, de véritables trésors informationnels quand elles sont qualitativement contrôlées et remises à jour. C'est aussi la *start-up* Linkurious⁷ et son *web service* d'exploration dynamique d'informations très contextualisées comme les données téléphoniques dans la résolution d'affaires criminelles, les réseaux de

5. http://www.lemonde.fr/international/visuel/2010/12/06/wikileaks-lire-les-memos-diplomatiques_1449709_3210.html.

6. <https://matlo.com/>.

7. <https://linkurio.us/>.

comptes bancaires ou les réseaux d'investisseurs de la base Crunchbase. Linkurious, au passage, équipe aujourd'hui les journalistes de l'ICIJ (l'International Consortium of Investigative Journalists⁸), un réseau international de 190 journalistes d'investigation issus notamment des grands journaux américains) dans leurs travaux autour du SwissLeaks. De plus jeunes encore dessinent en ce moment même le projet de cartographier les réseaux juridiques opaques du LuxLeaks (ou « Luxembourg Leaks ») où 343 sociétés multinationales ont mis en place un système d'optimisation fiscale au Luxembourg avec l'aide de cabinets spécialisés. Le terrain des documents juridiques représente à lui seul aujourd'hui un immense territoire vierge de toute exploration cartographique, une formidable terre d'aventure pour les amoureux des *data sciences*.

On comprend que les méthodes de cartographie des réseaux peuvent devenir des vecteurs de communication très polémiques et les graphes spatialisés des armes politiques redoutables lorsqu'elles sont diffusées dans l'espace public. Le principe d'accessibilité aux données permis par le Web (par exemple sous la forme d'applications coopérantes via des API) et des technologies ouvertes comme Gephi participent de la culture actuelle de l'analyse et du croisement des *data*. Selon moi, les premières pistes ouvertes par des organisations de la société civile, des *start-up*, des journalistes, des associations ou de simples citoyens ne constituent qu'une première étape modeste au regard des innombrables silos informationnels qui attendent d'être explorés, croisés, interprétés. Personnellement, je souhaite vivement que ce mouvement se développe comme le montre l'exemple suisse⁹. Ce travail de recueil et d'analyse de l'information peut d'ailleurs être entamé modestement, à la manière d'une pratique artisanale, comme en marge des *data sciences* très équipées. C'est le chemin emprunté par les journalistes suisses, en trois opérations majeures : rassembler, croiser, spatialiser.

Les lignes et les colonnes du fichier Excel, remplies soigneusement au fur à mesure de l'enquête, n'attendent plus que d'être croisées puis spatialisées, une fois passé un seuil quantitatif suffisant pour faire apparaître... l'organisation d'un système illégal. Et les journalistes suisses prennent soin de noter tous les détails de la métamorphose des *data* en carte : « Quand un de mes collaborateurs est venu dans mon bureau avec la liste des 55 mandats du projet Insieme délivrés de gré à gré, se souvient le vice-directeur de l'Office fédéral de la construction et de la logistique René Graf, j'ai tout de suite vu que nous avons un souci ».

8. <https://www.icij.org/>.

9. <http://dievolkswirtschaft.ch/fr/2015/06/2015-07-tanner-franz/>.

C'est de cette opération que naît la puissance de la cartographie d'information à base de graphes relationnels : la « carte » rassemble et concentre (comme un *atlas*), mais elle croise aussi et enrichit (comme un *filtre*) en distribuant des liens qualifiés entre les éléments de ce vaste système de corruption :

À chaque fois que la Confédération achète quelque chose, elle le fait pour 5,4 milliards de francs par an – elle se déplace dans un environnement où les relations entre fonctionnaires fédéraux et entreprises privées s'entremêlent. C'est ce que montre une analyse du *Matin Dimanche*. Les données de 1 745 mandats fédéraux publiés sur la plateforme officielle Simap¹⁰ entre 2009 et 2012 ont systématiquement été examinées en fonction des réseaux de relations. Cela a permis de mettre au jour une série de cas problématiques, dont certains relèvent clairement du copinage.

Et c'est toute la valeur de cette carte publiée dans les deux journaux suisses : les parcours inédits qu'elle propose s'inscrivent au niveau où sont croisées les informations, et non sur les « données brutes » qu'il a fallu construire au départ. Et les journalistes suisses poursuivent ainsi :

Le travail journalistique peut donc, une fois cette liste de correspondances établie, se porter sur l'étude de ces quelques cas particuliers, mettant au jour des relations d'affaires teintées de favoritisme ou de copinage. Ce travail sur les bases de données offre une vision que ni la Confédération, ni les entreprises, ni les autorités de contrôle externes n'étaient jusqu'alors en mesure d'apprécier.

Ce n'est donc pas seulement la possession des données qui importe, mais l'art de mettre en scène les réseaux qu'elles dessinent. Pour la gloire du cartographe... ou à ses risques et périls.

* * * * *

10. <https://www.simap.ch/>.

Nous touchons là les particularités des cartographies du « social », « individus » ou « organisations »¹¹. Car nous ne sommes plus seulement dans la prouesse technique ou l'art de la synthèse, mais aussi dans le champ politique des *indicateurs de gouvernance*, des technologies de *l'identité* (les noms propres sont mentionnés dans les cartes, et donc dans les journaux en Suisse) et l'analyse des formes organisationnelles (c'est l'ensemble d'un système d'ententes illicites qui est dénoncé). La cartographie d'entités nommées a donc quelque chose à voir avec la *traçabilité du social*. Rendre public l'ensemble d'une structure à travers les schèmes d'interaction d'un ensemble d'entités, la révéler à la lumière d'une carte peut ainsi constituer une arme redoutable, et difficile à manier. Car il ne s'agit plus de cartographier des gènes, des sites web, des brevets ou des espèces animales dans un biotope, mais bien *nos contemporains eux-mêmes* en rassemblant et en croisant les informations qu'ils produisent (souvent malgré eux ou sans se soucier la plupart de temps du fait qu'elles peuvent être rendues publiques). C'est tout le principe du *social data mining* : les données, une fois considérées comme des *traces*, finissent par composer des *empreintes*. La cartographie d'entités sociales et de leurs relations mutuelles relève donc d'un pouvoir, exercé en premier lieu sur ceux qui figurent *dans* les cartes. Dans certains cas, ces cartes d'entités nommées sont mobilisées à titre d'indicateurs pour la décision ou la stratégie et il m'est arrivé bien des fois d'en produire dans le cadre d'une diffusion restreinte (notamment pour des entreprises). Mais, plus le cercle de leur diffusion grandit, plus l'exercice de ce pouvoir de révélation devient délicat à manier. À titre de scientifique, le cartographe peut développer ses procédures comme s'il s'agissait d'élargir ou d'enrichir l'observation d'un phénomène. Mais c'est toujours à titre d'*acteur* qu'il assume (plus ou moins) son pouvoir d'observation (ou plutôt la traçabilité) du social, lui-même pris dans le jeu complexe des phénomènes humains. Les journalistes suisses trouvent ainsi dans la cartographie un puissant moyen d'accompagner un discours d'analyse, mais aussi une arme dénonçant une réalité qu'ils contribuent à modifier ou à tenter de le faire. Le contenu de la cartographie comme le moment de sa publication ou les médias qui lui servent de vecteurs contribuent à transformer l'instrument en arme politique.

Entrons-nous dans un âge de la transparence ? Peut-être. Pourtant, ce qui explose ce 25 novembre 2012 en Suisse est encore relativement rare, mais nourrit, en tous les cas, cette curieuse appétence pour la visualisation des « réseaux d'influence »,

11. Les termes ne sont pas les bons dans notre perspective, pas plus que d'autres termes ou concepts mobilisés en sociologie comme « types », « catégories », « habitus », « classes » ou « champs ». La cartographie d'entités nommées est basée sur des graphes relationnels et donc sur l'étude de *patterns* d'interaction à grande échelle et à partir de données en masses. Les « objets » qu'elle vise à interroger suivent des logiques génératives de *consécution* (les interactions s'enchaînent entre différentes entités sociales) ou de *complémentarité* (les entités sociales interagissent à l'unisson ou en complémentarité). Je ne sais pas ce que vaut cette distinction, mais elle a le mérite d'insister sur l'étude du social comme dynamique de *l'enchaînement* plutôt que comme pertinence des *types*.

des « *lobbies* », des « groupes informels », des systèmes de propriétés et des prises de participation, voire bientôt des transactions financières, des appels d'offres de marchés publics et leurs attributions à différents types d'acteurs économiques. À l'extrême, la cartographie d'informations et d'entités sociales peut aussi alimenter l'obsession des complots généralisés, comme ces cartes des flux bancaires qui attestent de la complexité des transactions commerciales mondiales, mais non, à mon sens, d'une stratégie explicite de dissimulation. Tous ceux qui se lancent dans ce type de cartographie peuvent très vite s'apercevoir de leur pouvoir, mais aussi de ses limites légales, éthiques ou politiques. Ainsi, quelques mois avant la parution de l'étude suisse, il en est certains parmi l'équipe de jeunes ingénieurs qui m'accompagnaient qui ont contribué d'un point de vue technique et méthodologique à un projet initié par Greenpeace-France, FaceNuke¹². Lancée le 13 avril 2012, la cartographie interactive (en SigmaJS, la librairie développée par Alexis Jacomy) avait pour objectif de décrire le réseau de relations qui unissait la plupart des acteurs connus du « *lobby* français de nucléaire ». Très rapidement, la belle cartographie du réseau social fait place sur le site web de Greenpeace-France au message suivant :

12. <http://greenpeace.fr/facenuke/>.

Image 13.3

Cet outil est une cartographie des personnalités, des organisations, une visualisation du réseau social qui compose aujourd'hui la politique énergétique de notre pays. FaceNuke permet ainsi de visualiser les personnalités connues et moins connues, les enchevêtrements d'entreprises, d'associations... C'est la volonté de faire la transparence sur le fonctionnement du système énergétique et notamment nucléaire français qui a guidé cette démarche. Certaines des personnalités présentes dans la cartographie n'ont pas apprécié cette démarche de transparence et ont souhaité que soient retirés leurs noms. Pas parce que les informations étaient erronées, non. Parce que leur nom était utilisé sans autorisation. Plutôt que déséquilibrer le graphe de FaceNuke en retirant certains des noms les plus importants, car, au cœur même du système, nous avons choisi de mettre l'outil en suspens. Le temps, pour nous, de revenir avec de nouveaux éléments. Parce que nous pensons que, dans les mois à venir, la transparence sera plus que jamais nécessaire.

Source : FaceNuke, par Greenpeace France.

C'est le genre d'événement qui permet de mesurer le pouvoir sociétal de la cartographie d'information, et du cartographe. En un sens, ce type de *sociogramme* propulse la cartographie d'informations relationnelles au centre d'un débat où se mêlent les questions politiques (car il s'agit dans le cas de FaceNuke des puissants de ce monde), éthiques (peut-on tout dire ou dévoiler ?) ou stratégiques (quand communiquer ?). Personnellement, j'aime ces rares moments où l'on perçoit la carte comme un événement et non plus seulement comme un instrument, illustrant le savoir-faire artisanal du cartographe. Quand elles génèrent le « buzz » pour leur pouvoir de

synthèse et de dévoilement, les cartes à base de graphes retrouvent une place de premier plan alors qu'elles sont le plus souvent à peine tolérées en marge des techniques de *knowledge discovery* ou de *data analysis*, toujours « coincées » dans les organisations entre « service informatique » et « documentation » ou « veille ». Le projet FaceNuke comme la grande carte publiée en Suisse en décembre 2012 illustre parfaitement le rôle de *pivot* que joue la technique cartographique entre les *data* d'origine et les multiples interprétations possibles qui s'ensuivent. Les spécialistes du *data journalism* empruntent ainsi le même parcours que les expérimentateurs en *web mining* face aux *patterns* statistiques et visuels des graphes d'URL et de liens hypertextes. À une différence près, et de taille : entreprendre de cartographier le « social », surtout s'il s'agit de dévoiler les coulisses des pouvoirs en place, c'est inmanquablement s'exposer (ou se voir rappeler) le rôle politique du cartographe lui-même. En d'autres termes, on ne peut douter qu'en dévoilant les structures ou les méandres du pouvoir exercé par des acteurs clés (groupes ou acteurs individuels), la cartographie devient un instrument particulièrement efficace de transformation de ce même pouvoir.

Dans ces moments-là, le cartographe apparaît clairement comme un *maître du jeu*. Il contrôle, en premier lieu, toute la *machinerie sémiologique* que suppose une mise en scène de l'information. Il décide de l'éclairage appliqué aux données, comme un point de vue particulier, délibérément assumé. L'exercice de dévoilement démontre combien toutes les cartographies d'information relèvent d'une stratégie de communication, variable suivant les contextes et les objectifs poursuivis. Même dans les univers scientifiques les plus « aseptisés » où la recherche « d'objectivité » semble dominer, chaque choix importe en termes de traitement des données. On ne choisit jamais impunément les opérations d'extraction, de qualification, de croisement des *data* originelles qui, projetées sous forme de graphe spatialisé, font apparaître ce qui ne pouvait l'être sous forme de listes ou de tableau de données. Et c'est sur ce type de pouvoir que jouent les journalistes suisses quand ils rendent présents et attestables des connexions sociales et leurs *patterns* jusque-là implicites ou auparavant inconcevables dans leur forme (un réseau de connivences) comme dans leur contenu (les marchés publics). Et il faut dire que le champ du croisement des informations associées aux entités nommées, en masses accessibles sur le Web aujourd'hui, en est à ses balbutiements et que le potentiel est infini. Par exemple, certaines personnalités qui sont intervenues pour interdire la publication en ligne de FaceNuke sont par ailleurs présentes dans d'autres cartographies que les ingénieurs qui m'accompagnent

aiment à produire régulièrement (disons à titre « d'exercice pédagogique »). Il leur arrive ainsi de générer régulièrement le graphe des personnalités plurimandataires des entreprises du CAC 40 : en premier lieu, le graphe (bipartite) où chaque personnalité est reliée aux entreprises où elle siège dans le conseil d'administration, dessinant ainsi un réseau d'influence potentielle ou de partage de l'information alors que nous sommes censés évoluer dans une société de la concurrence. La loi française en autorisant jusqu'à 6 mandats occupés simultanément dans des entreprises du CAC 40, on peut alors extraire des données (issues de Wikipedia) une vue intéressante des réseaux qui se dessinent à grande échelle, comme celui des réseaux d'entreprises du CAC 40 reliées entre elles par différentes densités de « partage » d'administrateurs communs.

Image 13.4

Enquête sur les administrateurs du CAC 40.

Maître de la mise en scène sémiologique, le cartographe joue aussi du pouvoir social que lui confère sa capacité à agréger des traces numériques qu'il transforme en empreintes d'entités sociales, presque en portrait. Dans ce travail, on ne saurait décider s'il s'agit d'une sociographie qui vise la *description* ou d'un travail engagé politiquement ou éthiquement et qui, lui, vise la *dénonciation*. Je sais, la frontière est mince, surtout si l'on exerce dans le champ des sciences humaines et sociales où les deux se confondent largement. Mais la carte que nous livre le cartographe fonctionne comme un *piège* de la même façon que Louis Marin décrivait l'appareillage rhétorique du récit¹³. Devant la cartographie des réseaux d'influences potentielles du CAC 40, comme dans celle publiée en 2012 par *Le Matin* en Suisse, chacun *doit* s'interroger sur l'interprétation de ces *patterns* en termes de « concurrence » fictive ou de réseaux d'entente illégaux. Tout comme chacun *doit* s'interroger sur la question du partage de secrets des affaires dans l'univers des grandes sociétés du CAC 40. Mais, surtout, on devine combien le cartographe pourrait aller plus loin pour obtenir d'autres empreintes d'entités sociales et/ou leurs *patterns* d'interaction, selon les mêmes principes de croisements des données, mais livrant des portraits beaucoup plus riches, plus précis. Par exemple, les données de cette « carte du CAC 40 » pourraient être croisées avec celles livrées par les biographies des plurimandataires (elles aussi disponibles sur Wikipedia pour la plupart) ou encore les documents juridiques dans lesquels ces personnalités sont partie prenante. Évidemment, la loi, ou l'éthique, en interdirait partiellement le principe. En un sens, cela soulignerait un peu plus le pouvoir technique de la cartographie, et son potentiel de subversion.

La cartographie d'information peut donc être mobilisée comme un instrument de stratégie et d'action, *instrumentalisée* dans un processus de décision. Ce pouvoir de dévoilement de *patterns* d'interactions sociales qu'elle détient a récemment suscité l'intérêt de différents types d'acteurs occupant, de différentes façons, des responsabilités en termes de gestion des compétences et des organisations. C'est le cas notamment des services des ressources humaines de grands groupes industriels : à partir des données d'inventeurs dans les brevets, de comptes de réseaux sociaux et/ou de la documentation interne des entreprises, il est possible d'observer précisément la structure et l'intensité des réseaux d'innovation technologique en R&D. De les observer... et de les évaluer, les orienter ou les réorganiser. On peut aussi cartographier les activités d'un concurrent... Ces terrains d'expérimentation restent encore discrets, et toujours très sensibles. En entreprises comme dans les institutions publiques, la production de cartographies d'entités nommées génère la plupart du temps les mêmes réactions. J'ai souvent observé combien les spectateurs suivent un cheminement intellectuel très précis dans leurs premières appréciations d'un travail cartographique : le « Comment c'est fait ? » s'efface après quelques

13. L. Marin, *Le récit est un piège*, Paris, Minuit, 1978.

explications techniques pour faire place au (fameux) « Où suis-je ? Où sont mes collègues ? Où sont nos services et nos projets ? Où sont nos concurrents ? ». Le troisième temps, quand il advient, a toujours quelque chose d'une *réussite* pour le cartographe : « Comment utiliser cette carte dans mes activités ou mes projets ? » ou, mieux encore, « Comment modifier ma position, pour devenir plus *central*, plus *relié* à mon environnement, *couvrir* l'ensemble du domaine ? ». Je me revois discuter, parfois âprement, des choix techniques jusqu'au moindre détail, sans oublier les fois où l'on me demanda d'anonymiser une carte dans le cadre d'une démonstration publique ou d'adopter une autre technique de spatialisation des données qui serait moins défavorable à tel ou tel acteur ou organisation. Indéniablement, la cartographie est un exercice techniquement difficile, mais aussi socialement, et frise parfois le numéro d'équilibriste.

J'accepte souvent de réaliser ce type de cartographie, et pour deux raisons. La première paraît évidente d'un point de vue technique : même lorsqu'elles sont données sous forme d'ID anonymes (identifiants dans une base), les entités nommées constituent un principe simple et puissant d'agrégation et de croisement des informations. Cette « facilité » technique s'explique par le développement exponentiel du *big data* très orienté par l'exploitation des données personnelles, notamment sur les plateformes sociales ou les sites de e-commerce, sans parler de la sécurité. Si l'on ajoute à cela le développement du marché des données privées ou le rachat des *start-up* qui en produisent par de grands groupes, on comprend l'intérêt pour les grands graphes d'entités nommées. Car, après tout, ce sont les hommes et leurs activités que ces immenses stocks de données servent à décrire, à observer... ou à orienter. Les entités sociales sont donc à la fois des *sources* (productrices de traces) et des *cibles* (commerciales, politiques, citoyennes, actives...) dans un mouvement global où s'agrègent les informations temporelles, géographiques ou l'immense *data lake* des données d'usage. Imaginez ce que pourront produire demain comme *data* nos objets connectés...

L'autre raison tient dans ce que peut apporter la cartographie d'entités sociales dans ce qu'il est convenu d'appeler aujourd'hui la « conduite du changement » ou la transformation progressive des hommes et des métiers dans les organisations (dont les entreprises). C'est un autre rôle du cartographe comme acteur social, mais le changement qu'il vise ne s'effectue pas sur le mode de la révélation ou de la dénonciation. Il m'est arrivé de nombreuses fois de constater, souvent avec surprise, que les

cartes d'acteurs, de projets communs, de relations de partenariat ou de collaborations personnelles peuvent accompagner (ou même susciter) une réflexion générale sur des questions d'organisation, qu'il s'agisse de la conserver, de la modifier ou de la transformer. En révélant, la cartographie peut aussi servir de levier. Parmi les nombreuses anecdotes qui illustrent ce pouvoir de transformation des organisations, l'une concerne une fameuse école dont la direction voulait « rationaliser » l'organisation de ses cinq équipes de recherche en les fusionnant en partie pour en constituer trois seulement. Suite à la demande de la direction de l'établissement, nous avons produit une cartographie où figurait l'ensemble des chercheurs reliés aux différents projets financés sur les dix dernières années. L'examen collectif de la carte imprimée en grand format laissa l'équipe de direction quelque peu perplexe : cette carte faisait clairement apparaître *quatre* communautés de chercheurs et non pas les cinq laboratoires existants, ni les trois qu'aurait voulu constituer la direction. Si l'on tenait compte de la dynamique de coopération spontanée des chercheurs, notre chiffre indiquait clairement qu'il fallait revoir la façon de distribuer les chercheurs dans de nouvelles équipes. Mais plus encore : trois chercheurs parmi les 250 de l'école semblaient occuper des positions de « relais » entre les quatre communautés actives de recherche, ce qu'en théorie des graphes on désigne par centralité (plus précisément le *betweenness centrality*). Ils ne figuraient pas parmi les plus reconnus de l'école, et ne « rayonnaient » probablement pas en termes de reconnaissance académique. Or, visiblement, ils participaient régulièrement aux projets de recherche, mais chaque fois différemment, en se connectant aux quatre communautés détectées dans notre méthode (cinq ou trois pour la direction) de façon presque alternative. Enlevés du graphe (ce que je fis plus tard avec Gephi), ils laissaient apparaître un univers très cloisonné avec des « chemins » (ou un diamètre) très longs d'un bout à l'autre de la carte. Nous avons donc affaire à trois éléments « pollinisateurs », des « passeurs de concepts » ou simplement des « médiateurs » sans lesquels beaucoup moins de relations transversales existeraient entre les équipes. Apparemment peu spécialisés dans un domaine particulier de la recherche en ingénierie, ils jouaient pourtant un rôle clé en termes de structuration de l'espace d'interaction scientifique de l'établissement. Je laissai à l'équipe de direction, plongée dans une certaine perplexité.

Cet impact à long terme, presque fortuit au départ, a été le point de départ d'une réflexion globale sur les conditions de « félicité sociale » d'un projet cartographique *pour* et *dans* les organisations dont l'une des conditions repose clairement sur sa dimension collective et partagée. Entre la révélation publique (comme celle des journaux suisses) et les cartes confidentielles calculées pour les directions et la stratégie, il existe une autre voie, un type collectif de révélation, une expérience unique où une organisation semble parfois se révéler à elle-même. Si l'on prend le temps d'expliquer publiquement les techniques, de discuter des types de calculs appliqués

et de décider des types d'indicateurs sélectionnés pour construire les graphes (de collaborations, de partenariats, d'activités et de projets...), alors la cartographie se transforme petit à petit en *miroir* où chacun peut se voir, mais aussi, relativement, l'ensemble des autres acteurs. Comme une révélation intime, et pourtant collégiale. Dans ce cas, la cartographie d'entités nommées peut devenir un instrument précieux de stratégie collective, notamment en termes d'évaluation (d'un acteur, d'un projet, d'une recherche...) qui *ne peut plus* s'effectuer individuellement. De telles cartographies ont le mérite, le plus souvent, de mettre en exergue les foyers d'interactions intenses (par exemple dans certains projets innovants) et des espaces vides, ou peu développés. C'est toute la différence avec l'organigramme censé identifier et organiser verticalement les services et les métiers. Dans tous les cas, les cartes d'entités sociales et de leurs interactions mutuelles dans une organisation (avec les qualités et les défauts qu'elles soulignent) pointent de façon magistrale les fonctions de *management* (existantes ou défaillantes), depuis la gestion des ressources humaines jusqu'aux directions fonctionnelles. Là réside leur véritable potentiel de subversion.

De là, les pouvoirs des représentations de réseaux, si elles sont fondées sur l'exploitation de données d'usage, notamment en entreprise. Sur ce point, je suis toujours aussi surpris lorsque l'on me demande à propos d'une carte « Ce sont nos données ? », illustrant le manque ou la faiblesse de l'exploitation des jeux de données potentiellement accessibles depuis les fichiers clients, les catalogues produits ou la mise en place de projets transversaux en interne. Mais, une fois dépassée la question, il est symptomatique de voir les acteurs s'emparer petit à petit des cartes comme des espaces à occuper, à modifier éventuellement, à renforcer ou à abandonner... En somme, à conquérir. Cet effet de réalité ainsi produit me laisse entrevoir (mais très faiblement) l'importance des enjeux idéologiques de la cartographie d'entités nommées qui semble annoncer, comme au moment de la Renaissance, que la maîtrise des territoires (éventuellement sociaux) reposera aussi peut-être sur le système symbolique et culturel qui permet de les cartographier. Évidemment, les interactions sociales n'ont rien de commun avec les territoires physiques à titre d'objets, mais, en rendant présents sur une surface des mécanismes aussi abstraits que ceux liés à des interactions sociales, avec le jeu méthodologique, technique, graphique proposé par le cartographe, voilà d'un coup l'avènement d'un code partagé, de territoires à cultiver et d'une carte pour s'orienter. Des moments rares dans les organisations, et une forme d'expérience où le cartographe saisit de façon privilégiée que révéler, c'est là aussi être le premier.

Chapitre 14. La cartographie d'information ou *l'alchimie qualitative-quantitative*

« “Puisque vous avez désobéi au commandement que Dieu vous fit, allez parcourir une vie de peines et de fatigues”. Adam, que Dieu avait choisi pour notre père à tous, qu’il avait si noblement doué, reconnut sa faute et quitta l’idée de sciences pour en revenir au travail des mains qui fait vivre. Il prit la bêche, et Ève commença à filer. Plusieurs arts nés du besoin suivirent, tous différents l’un de l’autre. Celui-ci entraînant plus de science que celui-là, ils ne pouvaient tous être égaux ; car *la science* est le plus noble. Après elle, en vient un qui lui doit son origine et la suit de près, il vient de la science et se forme par l’opération des mains. C’est un art que l’on désigne par le mot de *peindre* : il demande la fantaisie et l’habileté des mains ; il veut trouver des choses nouvelles cachées sous les formes connues de la nature, et les exprimer avec la main *de manière à faire croire que ce qui n’est pas, soit* », Cennino Cennini, *Il Libro dell’arte*, 1437.

La cartographie d'information semble tout avoir de l'art de l'illusion. On la tient parfois en suspicion dans le milieu scientifique comme un objet déporté, relatif ou contingent parce qu'elle relève du champ de l'image ou du *design* graphique et non du « concept » articulé, autrement dit des « mots ». Ce qui fait la force visuelle et esthétique de la cartographie dans la mise en scène des *data* semble aussi faire sa faiblesse comme instrument cognitif où, souvent, on ne la tolère que sous la forme de la « visualisation » censée « traduire les données » (la *dataviz* et ses mirages) ou « accompagner » le discours, la pensée ou la décision. En somme, un jeu d'illusions visuelles comme on aime à définir les images au pays de Descartes. Telles que je me remémore nos expéditions cartographiques au cours de toutes ces années, les points de vue que nous avons produits sur les différents *data sets* pourraient effectivement sembler relever massivement des « trucs et astuces » dignes de l'art de la mise en scène, *par magie*. Comme dans un théâtre vitruvien où les événements vont apparaître dans un cadre réglé par l'illusion perspective, nos variables graphiques comme les liens courbes ou la distribution des couleurs pour créer des zones (les fameux *clusters* que nous aimions qualifier de « patatoïdes » visuels) et nos filtres de spatia-lisation avec les choix imposés par l'utilisation systématique des algorithmes de type *force directed* (dont le Force Atlas 2 évidemment) ont constitué les rouages nécessaires à la manifestation de ce que nous croyons être des *patterns* dans les *data*. Et,

au-delà, certaines propriétés que nous sommes parvenus à identifier à propos du Web en dépendaient aussi.

Pourtant, l'illusion cartographique qui aboutit *ici* sur un poster, *là* sur une interface numérique, repose sur une machinerie logique qui fait résonner tout l'appareillage conceptuel et technique de la construction de l'objet de science. La cartographie d'information, qu'on la pratique soi-même ou qu'on l'observe, apparaît aussi comme une *machine logique productrice de topologies interprétables*. Y compris en dehors de la question des réseaux, elle recèle des propriétés particulières, *logiques* si l'on veut, qui la rangent au côté de tous les autres instruments d'investigation et d'analyse, donc de *cognition*. À y regarder de près, l'appareil cartographique qui me permet de prendre des vues sur les *data* se compose d'une série de filtres, dont une partie est destinée à préparer l'espace graphique pour le déploiement d'une interprétation. Il ne s'agit donc pas que d'une mise en scène des *data* au sens d'une forme de *rhétorique visuelle*, mais aussi d'un traitement logique où s'équilibrent des processus de traitement de l'information tant *qualitatifs* que *quantitatifs*. En position de pivot entre les données et les hypothèses (ou l'interprétation), la cartographie constitue donc un *lieu* d'observation, ou un *moment*, très éclairant de ces deux mécanismes logiques par lesquels nous saisissons les objets de science, alors observables. C'est en ce sens que j'aime à définir la cartographie d'information comme une *machine logique*.

La problématique du « quantitatif » et du « qualitatif » s'inscrit dans cet espace de la pratique scientifique où nous construisons les phénomènes étudiés, à mi-chemin entre l'hypothèse qui le surplombe et les *data*, brutes, sur lesquelles ils émergent de façon logique et organisée. En ce sens, elle n'est pas spécifiquement liée à la sociologie, à l'anthropologie, ni même aux sciences humaines et sociales : elle renvoie à ce jeu où s'associent tant des *qualités* que des *quantités* dans un modèle d'intelligibilité des données, et un seulement parmi d'autres possibles quand on adopte les procédures d'un champ de recherche comme je le fais avec l'approche réseau. Ce qui se joue localement en ingénierie des données et des connaissances sous forme de curseurs de saisie d'un phénomène relève d'une problématique plus générale en science, et au-delà. On peut réduire à grands traits les principes des orientations qualitatives ou quantitatives qui me semblent représenter deux façons complémentaires de construire l'objet de science, tel que l'on peut l'observer ou l'analyser et dont la cartographie est une forme technique. Spontanément, on peut concevoir *l'orientation quantitative* comme l'univers des mesures et des variables, de la recherche de *patterns* (sociaux par exemple) indirectement observables, des effets de masse et de seuils statistiques. Le travail scientifique sur des hypothèses quantitativement argumentées peut être d'une grande complexité dans la manipulation des opérations mathématiques, comme la recherche des « axes » ou des « composants » dans des

univers de données multidimensionnelles. Elle me semble fondamentalement liée à la quête de régularités, voire de lois accessibles une fois seulement dépassés certains seuils dans les masses de données. Ce qui ne semble pas poser de problème méthodologique ou critique en physique ou en biologie se révèle particulièrement discuté en sciences humaines et sociales où l'usage des données chiffrées et en nombre serait sujet à une forme inhérente de « relativité » (du chercheur, des instruments, des cadres théoriques sous-jacents) et à des formes de « positivisme » dont les fondements théoriques ou idéologiques n'ont pas été suffisamment interrogés. On peut ainsi opposer à l'orientation quantitative, une *démarche qualitative* orientée vers les singularités, les interactions locales ou le suivi longitudinal d'un nombre restreint, mais richement décrit, d'acteurs ou de petits groupes. Contrairement aux supposées réductions quantitatives, le curseur qualitatif me semble gouverné par la recherche foisonnante des moyens de dilater le « phénomène » ou « l'objet social », par exemple en y participant soi-même comme en ethnométhodologie, en invoquant la complexité du fonctionnement humain avec ses dimensions culturelles ou psychanalytiques, ou alors (comme je l'ai observé en ergonomie cognitive) en consacrant beaucoup de temps et de moyens d'observation ou d'interprétation à quelques minutes d'activité d'un acteur devant une console d'ordinateur.

Oui, je sais : j'entre là dans les débats érudits de l'épistémologie, de l'anthropologie, des sciences de la cognition. Le débat entre démarche *qualitative* et démarche *quantitative*, à lui seul, occupe une place majeure en sciences humaines et sociales, surtout en sociologie et en sciences politiques. Il s'y joue des postures, des hypothèses et des méthodes a priori irréductibles, comme si la dichotomie *qualitatif-quantitatif* ouvrait sur deux champs distincts de science. Et les sciences exactes n'y échappent pas non plus, bien que le débat y soit moins sensible. Ce débat ancien trouve aujourd'hui un écho particulier, et très polémique, quand on envisage, de surcroît, les technologies numériques, les masses d'information et les puissances de calcul. Cette résurgence du débat a été alimentée récemment par certaines perspectives induites par le *big data*, en particulier dans la découverte et l'exploitation de tendances statistiques massives dans les données numériques aujourd'hui accessibles. Des provocations de Chris Anderson dans *Wired*¹ qui annonce la fin du modèle scientifique expérimental avec l'apparition des *data sciences* (« But faced with massive data, this approach to science –hypothesize, model, test– is becoming obsolete ») aux réactions conservatrices des tenants des modèles abstraits, théoriques et érudits en sciences humaines face à cette foule d'ingénieurs et leurs infrastructures de traitement à grande échelle, les postures des uns et des autres s'arc-boutent sur des distinctions que le cartographe que je suis a du mal à discerner dans la pratique. Ma

1. C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired*, 2008, <https://www.wired.com/2008/06/pb-theory/>.

perplexité face à ce débat redouble quand je m'aperçois des autres problématiques que la distinction qualitatif-quantitatif emporte avec elle dans son sillage comme celle du *manuel* et de *l'automatique* (dès qu'il y a des machines et des ingénieurs, certains s'imaginent qu'il ne s'agit que de traitement quantitatif de l'information) ou celle des rapports de l'hypothèse ou du *modèle théorique* avec les *données expérimentales* (l'effort de théorisation étant présenté comme un exercice qualitatif).

Quand bien même je voudrais échapper à ce débat généralisé, il m'est imposé de toutes les façons quand je parle de cartographie de l'information. En particulier, c'est systématiquement le cas lors des conférences où je présente mes activités : je ratte bien évidemment la cartographie de l'information (et la science des réseaux) à la culture actuelle des *data*, en termes de méthodes comme de « posture » (sous l'angle de l'*open data*, du *big data analytics*, du *data scientist*, du *data intelligence*, voire du *grid computing*, auxquels ont fait référence depuis quelque temps des revues comme *Nature*, *Science* ou *O'Reilly Radar*). La cartographie, après tout, est une façon de traiter des masses de données et les *network sciences* de les expliquer. Mais souvent, à peine ai-je eu le temps d'afficher un poster sur un mur, que tout un ensemble de questions récurrentes pleuvent et m'obligent à développer des débats sans fin sur le rôle des « machines » et des « statistiques » à l'heure des réseaux distribués d'information. Les procédures d'analyse automatisées de vastes ensembles de données modifient-elles le rôle des experts, voire leur place dans la construction des connaissances ? L'ingénierie quantitative en *data sciences*, avec leur démarche inductive, peut-elle se passer de modèles qualitativement construits ? La cartographie, avec ses techniques et ses méthodes, est assurément un instrument de réduction des masses, mais sa *relativité* (ce n'est qu'une méthode parmi d'autres) n'impose-t-elle pas une *réduction* des pistes de réflexion, comme une forme d'*appauvrissement* ? C'est dans ces moments que j'aimerais pouvoir transporter mes interlocuteurs dans mon atelier de cartographie et leur montrer, à travers les gestes et les procédures, que ce qu'ils considèrent comme des limites ne sont que des contraintes de départ pour accéder à des possibles dont, souvent, ils ne soupçonnent pas la rationalité et la dynamique.

Les débats gagneraient en pertinence si l'on s'attachait quelques instants à la description des traits saillants de cette ingénierie des données, du moins tels qu'ils peuvent m'apparaître, et où la cartographie d'information tient sa place. Cela permettrait peut-être d'éviter les caricatures que l'on voudrait parfois dresser à

propos de l'ingénierie actuelle des *data* supposée *quantitative* (le culte des machines, l'obsession des statistiques, l'absence de recul critique sur les pratiques et les méthodes). Pour apporter un peu de clarté au débat, il me semble que l'exercice de la cartographie d'information constitue un excellent terrain d'interrogation de ces problématiques dont la dichotomie *qualitatif-quantitatif* semble être le nœud central. Quand on regarde de près une chaîne de traitement de l'information pour produire des cartographies, les opérations manuelles et artisanales menées sur les données (depuis l'extraction jusqu'aux *patterns* visuels cartographiés) mettent à jour de façon explicite l'*imbrication* des procédures tant qualitatives que quantitatives que l'on fait porter sur les données. Dans l'univers de l'artisanat de haute technologie dans lequel je baigne, la mise au point de prototypes oblige à décomposer des dimensions et des procédures qu'un algorithme industrialisé masquerait sous la forme d'une « boîte noire », éventuellement propriétaire et cadenassée par un brevet. L'observation du travail minutieux du cartographe dans son atelier éclaire le processus de croisement systématique du qualitatif et du quantitatif dont dépendra, au final, la nature du phénomène scientifique visé. Ce sont les deux dimensions constitutives de l'information dans le milieu dans lequel je travaille. De la simple feuille de papier avec ses listes et ses tableaux à la table Excel et ses fonctionnalités de calcul jusqu'aux systèmes « temps réel » sur le Web, ces deux formalismes logiques imposent leurs dimensions à nos instruments techniques de traitement au rang desquels figure la cartographie d'informations.

Si on la considère sous l'angle d'une *machine logique*, apparaissent alors les deux curseurs quantitatif et qualitatif qui dessinent une fenêtre de saisie de phénomènes décrits avec des informations. Comme avec un métier à tisser, les motifs générés dépendent de formes avancées d'hybridation de qualités (couleurs, textures...) et de quantités (sur combien de lignes ?). En ce sens, la cartographie d'information, comme les autres instruments scientifiques, impose sa grille de saisie aux « objets » visés. Celle-ci s'applique aussi bien à des gènes, des publications scientifiques, des organisations sociales, des molécules, des abonnés, des concepts qu'à des « acteurs » ou des « interactions ». Dans la pratique, le curseur qualitatif et le curseur quantitatif doivent être réglés avec soin. C'est un moment important pour un cartographe qui doit au final associer les propriétés de l'espace des données prises en compte avec les propriétés du plan graphique et ses variables. La pertinence de la cartographie d'information comme instrument d'investigation n'est effective que dans certaines limites liées quantitativement aux masses d'informations et qualitativement au nombre de dimensions. Quantitativement, il ne faut ni *trop*, ni *trop peu* d'éléments ou de « lignes » pour pouvoir régler manuellement là aussi les différentes échelles de l'information. Dans le cadre des graphes, un nombre trop faible de nœuds ne fait pas un réseau, qui plus est si le nombre de liens est également réduit. Les *patterns*

organisationnels recherchés (la topologie d'un système) n'apparaissent qu'une fois passé un certain seuil de données. Qualitativement, il ne faut ni *trop*, ni *trop peu* de dimensions dans les données pour pouvoir manuellement jouer sur leur corrélation (jeu manuel dans le cadre de la cartographie d'information relationnelle telle que je la pratique, mais on peut en faire un jeu très abstrait et/ou automatisé dans ce que l'on désigne par méthodes d'analyse multidimensionnelle). Ces réglages simultanés des deux curseurs de saisie de l'instrument conditionnent la pertinence du cliché cartographique : a-t-on suffisamment réduit les masses pour faire apparaître un *pattern* identifiable ? A-t-on identifié suffisamment d'identités remarquables (ou de dimensions associées) pour les voir apparaître dans l'ensemble des masses ?

Image 14.1

Cartographie pour un laboratoire de recherche. Ici ont été croisées les données « auteurs » et les données « publications » parmi les nombreuses corrélation possibles dans une table de données. Cette carte intègre aussi un croisement des « publications » avec les « dates de parution » perceptibles à travers les différentes colorations de bleu attribuées aux publications.

Une cartographie de ce type implique un double travail de manipulation d'une table de données qui peut s'avérer d'une grande complexité.

Les mécanismes sur lesquels repose le cliché cartographique supposent l'exploitation d'une structure de données dont la table est l'archétype. Au sens large, elle peut être réduite à une « simple » liste (un fichier texte utilisant la convention CSV) ou élargie à une structure de base de données (éventuellement distante et interrogeable via une API). Dans tous les cas, la production de cartographie dépend d'une combinatoire de saisie (interrogation, requêtage, extraction, transformation...) des informations contenues dans les tables de données. Encore une fois, il convient de souligner que cette combinatoire ne relève pas seulement du choix ou de la posture théorique du chercheur (souvent déterminante en sciences humaines et sociales), mais aussi

de la façon dont techniquement nous construisons les objets étudiés à travers des modèles de données. Ainsi, tout commence par ce *prêt à découper-composer* qu'est le fichier XLS, disons une structure de table de données.

Image 14.2

Noeuds	Types	Liens	Coordonnées GPS		Date
P1	Projet				2006
P2	Projet				2007
MC1	Mot-clé	P1, P2			
A1	Acteur	P1	48.856614	2.3522219	
A2	Acteur	P37	47.218371	-1.553621	

Structure de table des données.

Combien, quantitativement, de lignes (d'objets à compter), combien, qualitativement, de colonnes (de descripteurs, d'angles sous lesquels les éclairer ?) ? La structure de données livre donc deux curseurs à manipuler pour construire l'objet d'observation : la fenêtre des lignes dont plus le nombre est grand plus le potentiel de détection de *patterns* est important selon les échelles ; la fenêtre des colonnes qui fixe un potentiel de corrélation des dimensions dans les données. La production de structures de graphes constitue une façon particulière de « prendre les données » en croisant deux colonnes parmi les autres qui partagent des objets ou des propriétés communes (principe de redondance de l'information ou de récurrence). C'est là toute la dynamique de la cartographie à base de graphe. Dans le cas des bases de notices bibliographiques, les publications peuvent être reliées entre elles par des liens de citation (directement accessibles depuis l'une des colonnes du tableau) ou bien par des liens calculés dans la table (présence de descripteurs communs, mêmes auteurs, mêmes laboratoires). Si les fenêtres de saisie de l'objet sont trop massives (du point de vue cognitif comme du point de vue des instruments), des outils de (pré)traitement automatique et de réduction sont nécessaires pour identifier des *clusters*, analyser des composantes multiples, agréger des dimensions, visualiser des croisements ou des résultats de seuillage. Les données imposent leurs modèles que l'on peut, en retour, façonner.

La double saisie des informations mobilisées n'est pourtant qu'un point de départ, le premier mouvement de la *machine logique*. Les lignes et les colonnes ouvrent potentiellement sur un jeu infini de croisements, de comparaison, de corrélation. En termes

de compétences humaines, il s'agit bien d'intelligence des données (*data intelligence*) et, à regarder travailler les ingénieurs en systèmes d'information, on se rend compte du degré d'*intelligence embarquée* que peut contenir un algorithme, lui-même intégré dans un processus ou une bibliothèque plus vaste. Là où les sciences humaines et sociales semblent avoir en partie figé la distinction entre les deux processus de production et d'analyse de l'information, l'ingénierie des *data* en exploite le potentiel, si possible *tout* le potentiel en tentant d'automatiser les corrélations comme dans le *big data* avec l'idée de chercher à identifier dans de grandes masses de données un ou plusieurs « traits » communs, révélant un *pattern* statistique qui n'est qu'une façon parmi d'autres d'isoler des identités partielles et partagées. Cette méthode *pattern based* peut être qualifiée de quanti-qualitative. Mais la méthode inverse, quali-quantitative, n'est pas incompatible, elle est même souhaitable : une propriété repérée localement (par exemple une information particulière associée à un nœud dans un graphe) est-elle repérable aussi à grande échelle ? C'est l'agilité avec laquelle on manie les deux opérations qui détermine souvent le nombre et la richesse des prises que l'on se donne sur les corpus de données numériques. Cette intrication des deux approches est pour ainsi dire native des recherches opérationnelles de type *data analytics* : dans les domaines de l'analyse des flux (par exemple les connexions quotidiennes sur réseau de téléphonie mobile ou les informations liées aux « parcours patients » entre des unités médicales dans un hôpital), les objectifs consistent souvent soit à modéliser les flux principaux, soit à « détecter des événements » qualitativement identifiables-calculables (curieux, remarquables, inédits...). Cette recherche repose sur la production d'un ou plusieurs modèles qualitatifs du phénomène (combinaison de traits spécifiques, distribués dans une configuration probable) appliqués à différentes échelles des masses de données (les quantités de données réunies pour valider le modèle pouvant donc se trouver à leur tour mobilisées comme traits qualitatifs à un niveau supérieur d'intégration).

Pour le moment, la cartographie d'information reste encore en marge de l'univers des réseaux dynamiques de données et rares sont les projets où l'on essaie de la produire automatiquement. C'est peut-être sa nature même : figer le processus de traitement, au moment propice d'un cliché pertinent. L'image me paraît appropriée si je calcule le nombre d'essais nécessaires pour caler la prise de vue en alignant précisément toute une série de filtres ! Dans son atelier, le cartographe que je suis façonne encore artisanalement sa chaîne de traitement de l'information, qui n'est jamais exactement la même d'un chantier à l'autre. Mais l'artisanat de haute technologie comme la cartographie d'informations a des vertus, disons, « pédagogiques », car on peut y décrire soigneusement, *au ralenti*, ce que l'automatisation cache définitivement : le cartographe qui règle ses curseurs et aligne différentes procédures de construction qualitative et quantitative de l'information. Et c'est maintenant dans la dynamique d'hybridation des deux procédures que prend forme la carte.

Image 14.3

La cartographie de l'information comme une boussole pour s'orienter.

La carte, presque achevée, porte un scénario exploratoire calqué sur les grandes directions d'analyse de données. L'exploration de la carte, donc de certaines propriétés des données du corpus, est réglée par la façon dont qualités et quantités *se contrôlent mutuellement*. La carte permet en premier lieu de faire apparaître certaines propriétés *qualitatives*, qui peuvent être vérifiées globalement ou localement en fonction de la façon dont elles *résonnent* dans les masses quantitatives. C'est le principe qui permet de classer les éléments d'un système, d'un point de vue ou d'un autre. Dans l'exemple choisi, il s'agit d'une cartographie de brevets fondée sur les liens de citation au sein du corpus. Je peux faire apparaître des propriétés globales, des *saillances* indiquant l'ossature générale du jeu de données : les brevets les plus importants en grosse taille – le plus souvent en couleur grise – ainsi que les principaux flux de citation. Je peux aussi « *zoomer* » sur une zone particulière du graphe et y vérifier des propriétés locales, valables seulement en contexte. Ce qui fait donc la variation qualitative de généralité ou de particularité des propriétés de corpus dépend de la façon dont elle est quantitativement vérifiée ou contrôlée dans les masses.

On pourrait en illustrer le principe de nombreuses façons. Par exemple, certains vont se concentrer plutôt sur les opérations d'identification, de sélection, de *ranking* et recomposer petit à petit des quantités (représentatives). C'est le principe de la recherche des types regroupant différentes identités (la notion de « profil » en *social data mining*) et les classant de plusieurs manières selon des critères qualitatifs (âge croissant des internautes, distribution de leurs liens affinitaires, métiers...). Mais c'est dans sa confrontation avec les quantités (ou les « masses ») que ce travail sur les critères qualitatifs se trouve validé – en réalité *contrôlé* : le *type* construit est-il suffisamment récurrent ? Mon ou mes critère(s) de sélection suffisent-ils à classer tous les objets de mon système (comme avec les algorithmes de classification automatique) ? Sur le Web, parmi les blogueurs, les « communautés voisines » de celle que j'étudie font-elles partie ou non de mon « corpus » ? Sélectionner (des traits), choisir (des types ou des familles), découper ou segmenter (le continu des masses), autant d'opérations qui guident la maîtrise des quantités. On pourrait illustrer le principe de mille et une façons, mais il paraît d'autant plus prégnant (et observable) quand il s'agit d'extraire des données d'un espace comme le Web qui ne livre pas de lui-même des principes clairs de découpage. Certaines opérations en *web mining* illustrent bien toute cette mécanique de la sélection, avec ses essais, ses erreurs ou ses incertitudes, dans le réglage par exemple du « focus sémantique » d'un *crawler* web (est-ce que je choisis les « bons » termes pour sélectionner les pages pertinentes ?) ou bien quand, avec le Navicrawler² (eh oui, je l'utilise encore !), j'archive à la fois un corpus de pages web, et aussi celles que j'ai exclues. Les « masses de données » font rêver pour leur grandeur estimée, mais il me semble qu'elles ne deviennent « quantités » que lorsqu'on leur applique des seuils, que l'on y repère des saillances statistiques, des effets de redondance ou qu'on les considère comme « représentatives ». En somme, qu'elles constituent un moyen de contrôle ou de validation de ce qui est défini ou sélectionné qualitativement. On réduit souvent le travail sur les quantités à l'usage des outils statistiques et mathématiques, conduit par des ingénieurs focalisés sur les grandeurs. C'est vite oublier, comme en *network sciences*, que les résultats de calculs statistiques (par exemple, le *small world phenomenon* et une distribution des connexions en loi de puissance) ne sont que des signatures, des indices qui ouvrent sur la recherche d'un principe d'ordre qui gouvernerait des réseaux ouverts, de taille gigantesque et dynamiques dans le temps. Encore faut-il mobiliser ou maîtriser un nombre suffisamment important de « quantités » pour faire émerger des propriétés ou, à l'autre extrême, que l'on dispose bien des algorithmes et de la puissance de calcul nécessaires à leur exploitation. Entre les deux, tous les niveaux de « *zoom* » sont possibles, du plus petit élément singulier aux effets de structure, à condition de spécifier que les « masses » ont été contrôlées comme des « quantités » calculables, des ensembles composés d'unités dénombrables.

2. <http://webatlas.fr/wp/navicrawler/>.

Mais la carte permet, parallèlement, de faire apparaître certaines propriétés *quantitatives* qui peuvent être vérifiées quand on décompose un système en *composants* ou en *clusters* (a-t-on distribué tous les éléments dans des groupes homogènes ?) et, au-delà, quand on cherche à recomposer le corpus selon une logique de complémentarité, comme pour en dessiner le plan global ou l'*organisation*. Notre capacité à décomposer/recomposer selon des frontières dans les masses dépend des critères adoptés. Dans l'exemple choisi, la distinction des *composants* ou des *clusters* effectuée repose sur les variations de distribution des liens de citation, auxquelles sont très sensibles certains algorithmes de détection de communautés sociales comme Modularity. De même, ces derniers s'assemblent en une organisation spécifique selon un ou plusieurs critères de complémentarité. Ainsi, dans les cartographies à base de graphes relationnels, la restitution d'un modèle d'intelligibilité du corpus passe par l'identification de sous-ensembles constitutifs, ses « régions » si l'on veut, en termes de cartographie, et par la compréhension de la ou des logiques qui préside(nt) à leur distribution. Mais le découpage des masses peut s'effectuer selon d'*autres* critères présents dans la table de données (ou que j'ai moi-même produits au cours de l'analyse) de façon à trouver un critère optimal de segmentation des masses. Par exemple, je peux produire avec le même jeu de notices de brevets une cartographie des liens entrants et/ou sortants, isolant les brevets les plus anciens et/ou les plus stratégiques (liens entrants qui s'accumulent au cours du temps) ou les brevets les plus récents (l'actuel « front de recherche »). Je peux aussi faire appel à d'autres dimensions du jeu de données comme les mots-clés communs entre les brevets, ou même les inventeurs. De là, il apparaît que notre capacité à découper quantitativement les masses d'information dépend de la façon dont nous faisons *tourner* le corpus sous différents angles en l'analysant sous différentes dimensions qualitatives, comme avec un prisme. Ce sont ces opérations successives qui font une partie du travail de *data scientist*. Et le travail peut être parfois long pour un cartographe, même accompagné par un dispositif de calcul : d'ailleurs, a-t-on jamais abordé un corpus de données à toutes les échelles et selon toutes les dimensions, d'un coup ? Peut-on même le faire ?

Cette maîtrise des deux curseurs logiques est essentielle dans le travail de conception d'une cartographie, qui est une forme de modèle statistique et visuel des données. Évidemment, il s'agit d'une pratique artisanale et très contextuelle, mais qui s'apparente selon moi au domaine plus général du *data analysis*, de la science des réseaux et, au-delà, des *computer sciences*. Certaines formules d'un atelier de cartographie,

maintes fois éprouvées manuellement, peuvent être en partie automatisées, voire intégralement. L'automatisation ne change rien fondamentalement à la grammaire ou aux formules quali-quantitatives ou quanti-qualitatives sur lesquelles repose la machine logique, seulement l'ampleur et la vitesse de traitement de l'information. L'automatisation permet le déploiement de procédures plus complexes, *embedded*, de mêler les types de traitement ou les algorithmes mobilisés, de croiser toujours plus vite les processus tant qualitatifs que quantitatifs de production de l'« information ». Ce qui fait débat du côté des sciences humaines et sociales, fonctionne donc « à plein régime », si j'ose dire, du côté des métiers du *data analysis* et de l'innovation technologique.

Dans ces domaines, les avancées expérimentales reposent sur l'exploitation continue de la dynamique du croisement des dimensions quantitatives et qualitatives, pour embrasser de plus vastes ensembles, pour peu que des formules pertinentes de traitement aient été figées dans un algorithme. Le terme de *formule* me paraît ici adéquat pour résumer le travail de conception d'une combinatoire complexe qui vise à isoler un type précis de propriété dans des données en faisant varier ou alterner processus qualitatifs et quantitatifs. Ainsi, de nombreux algorithmes mobilisés en *web mining* reposent sur des « recettes » ou des « formules » largement éprouvées, avec un « préréglage » des deux curseurs (seuils minimum et maximum de données web, dimensions qualitatives prédéfinies de l'analyse) : c'est le cas de HITS conçu par Jon Kleinberg qui opère à partir de différents degrés de corrélation entre liens hypertexte et détection de contenu permettant de découper des « masses d'informations » en ensembles cohérents. Certaines formules, d'ailleurs, « valent de l'or ». Elles sont au cœur des technologies numériques de l'information à l'image de ce qu'elles sont dans le domaine des cosmétiques et de la chimie fine. Souvent protégées par des brevets, certaines formules fonctionnent sous forme de « boîtes noires », propriétaires, volontairement inaccessibles. On prête même à certaines un pouvoir prédictif en identifiant les variables qui, parmi tant d'autres, semblent jouer un rôle majeur dans l'évolution d'un phénomène ou d'un système. D'autres, sont conçues comme des systèmes évolutifs, *auto-apprenants* et capables de modifier eux-mêmes l'alignement des filtres qu'ils intègrent. Au sens propre, les formules concentrent des formes artificielles d'intelligence.

Toutes ces formules algorithmiques contribuent à élargir de façon massive nos fenêtres qualitatives et quantitatives de saisie d'un phénomène. Beaucoup d'entre elles nourrissent aujourd'hui les applications les plus courantes et les univers quotidiens de manipulation de l'information. Sans être spécialiste de l'ingénierie *data*, chacun reconnaîtra l'étendue des formules de croisements complexes d'une simple application locale comme Excel, tout à fait étonnante, et l'on imagine ce que l'on

peut réaliser avec de vastes dispositifs de stockage et de traitement des données dans l'univers du *big data* actuel et de l'accès à des corpus distants, eux-mêmes déjà analysés et enrichis. La conception de formules ou de modèles de traitement de données concentre ainsi de nombreux efforts d'innovation, dans tous les secteurs et les métiers de l'information. En termes de services innovants, la formule peut être incarnée dans un dispositif né de l'agrégation originale d'une série de filtres analytiques qui portent sur une ou plusieurs dimensions des données et qui peut être concentrée dans une interface. En science, l'alignement d'une série de filtres, testé sur de nombreux *data sets*, peut être considéré comme un objectif en ingénierie de l'information et l'on ne compte plus les publications qui en détaillent la grammaire technique. En un mot, la formule préfigure le prototype (scientifique ou industriel) et fonctionne parfois en R&D comme une « boussole » pour l'orientation de l'innovation technologique et la mise en place de nouveaux services.

Dans son atelier, le cartographe d'information dispose désormais aujourd'hui d'une batterie étendue d'algorithmes ou de formules opérationnelles, qu'il combine en solutions hybrides, en les associant parfois à leur tour dans des ensembles plus vastes. Il m'arrive par exemple de faire appel aux compétences ou à la puissance de calcul d'une entreprise pour préparer un grand jeu de données. Une fois travaillées dans mon atelier quotidien, les données enrichies retournent parfois chez un autre partenaire, une autre entreprise qui va à son tour les passer au prisme de sa technologie... pour finalement être visualisées depuis ma machine. Ce que l'on désigne comme « intelligence des données » me semble relever massivement en phase exploratoire d'un processus de plus en plus étendu, géographiquement ou temporellement, incluant aussi bien des compétences humaines diverses que des applications distantes et coopérantes. L'art de la formule me paraît constituer le cœur de la démarche cartographique que j'aime à considérer comme une des formes de l'artisanat de haute technologie dédié aux *data*. Quand on le regarde précisément travailler dans son atelier, avec ses méthodes et ses instruments, le cartographe cherche seulement à comprendre la matrice des croisements ou à en éprouver le potentiel. De là, émergeront parfois quelques formules prometteuses d'où l'on pourra extraire un modèle de données plus vaste et plus robuste. Mais ailleurs ou plus tard, car, pour le moment, le cartographe reste plongé dans ses tentatives d'hybridation de filtres, faisant naître parfois un phénomène au hasard de ses manipulations combinatoires. Comme un alchimiste...

Chapitre 15. Cartographier la science

« *A map of science* is a spatial representation of how disciplines, fields, specialties, and individual papers or authors are related to one another as shown by their *physical proximity* and *relative locations*, analogous to the way geographic maps show the relationships of political or physical features on the Earth », Henry Small, “Visualizing Science by Citation Mapping”, *Journal of the American Society for Information Science*, 1999.

Lors d’une visite à l’École des mines de Paris, je passai voir la fameuse bibliothèque dont on m’avait maintes fois parlé. À l’évidence, le volume et la quiétude du lieu contrastaient avec les murs blancs des petites alvéoles du centre de recherche de mon université, et surtout l’effervescence qui régnait autour des machines lors de nos phases d’exploration du Web. Mais j’eus aussi le sentiment d’une sorte de retour vers tous ces livres imprimés qui avaient balisé mes études universitaires. C’est là que je retrouvai mes compagnons d’aventure¹ à qui j’avais proposé quelques semaines plus tôt de réaliser une carte des connaissances sur le thème du *véhicule autonome*². Le thème de notre recherche était à peine défini, mais suffisait pour réunir la petite équipe qui se composait d’un trio incontournable d’acteurs pour une expédition cartographique : l’expert du domaine pour la pertinence de son regard, le bibliothécaire pour sa maîtrise des sources (et l’ouverture à leur accès !) et le cartographe pour ses solutions techniques ou méthodologiques. Le sujet semblait passionner mes compagnons puisque nous étions au cœur de thématiques sensibles de recherche, aussi bien pour les chercheurs de l’École des mines de Paris que pour ceux de l’institut français des sciences et technologies des transports et de l’aménagement des Réseaux (IFSTTAR). Pour un néophyte comme moi, le thème évoquait surtout l’actualité du moment : la Google Car, les exploits de Tesla ou le *lapin lunaire* (le « lapin de jade ») que les Chinois voulaient expédier sur la lune quelques mois plus tard. La bibliothèque et ses annexes abritaient aussi des postes informatiques d’où il était possible d’extraire des catalogues des grands éditeurs de l’IST (Information scientifique et technique) comme le Web of Science (le WoS³) et, surtout, Scopus⁴ que je comptais exploiter particulièrement pour l’étendue de ses métadonnées et sa

1. Clément Boussard (chercheur à l’IFSTTAR), Akin Kazakci (chercheur à l’École des mines ParisTech), Doriane Ibarra et Jordane Raisin-Dadre de la bibliothèque de l’École des mines de Paris.

2. <https://ateliercartographie.files.wordpress.com/2013/10/rapportvehiculeautonome.pdf>.

3. <http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/web-of-science.html>.

4. <https://www.elsevier.com/solutions/scopus>.

réputation en termes de « couverture » des domaines techniques ou technologiques. Je suppose que la plupart des articles scientifiques que nous cherchions figuraient dans les revues imprimées et stockées quelque part dans la bibliothèque de l'école, mais dresser manuellement la carte des connaissances accumulées sur notre thème aurait pris des mois sans passer par les ressources permises par les catalogues éditeurs ou les systèmes internes de gestion des documents issus de l'IST. Ce que l'on a appelé pendant une période « l'informatisation des ressources documentaires » a autorisé toutes les formes de « recherche verticale » sur un auteur, une date, un mot-clé, mais aussi des opérations plus complexes, disons *latérales*, d'investigation, de comparaison, de tri croisé, de traitement automatisé des contenus et d'exploitation des métadonnées à différentes échelles. Le calcul a ainsi ouvert d'immenses possibilités d'exploration des masses de connaissances produites chaque année dans la recherche et la technologie, balayant tous les champs du savoir académique, de l'environnement aux nanoparticules, de la préhistoire aux recherches contemporaines sur les réseaux sociaux. C'est là, dans ce processus, que je comptais greffer les techniques de cartographie d'information aux structures de données des éditeurs, peut-être même les enrichir et, pourquoi pas, connecter d'une façon ou d'une autre l'univers des connaissances validées et formalisées à l'univers foisonnant du Web et de ses services. Je cherchais donc une façon très cartographique de faire dialoguer les deux univers...

J'entrais de nouveau dans le domaine de l'Information scientifique et technique que j'avais expérimentée durant un séjour à l'INIST. En effet, au cours de nos pérégrinations sur le Web, j'avais découvert qu'il existait *déjà* une communauté d'acteurs plongés dans la cartographie d'information, elle aussi à base de graphes et dédiée, non pas au Web, mais à « la science ». Enfin, disons les *produits finalisés* de la recherche scientifique nationale ou mondiale, et partout indexés minutieusement sous forme de « fiches » ou de « notices » : les publications scientifiques, les revues où elles sont publiées, les domaines auxquels elles appartiennent, mais aussi les brevets, les rapports techniques et tout ce qui a un rapport avec l'IST. L'univers de l'IST avec ses ingénieurs de recherche, ses documentalistes, ses bibliothécaires, et quelques rares chercheurs en sciences de l'information contrastait en bien des points avec l'univers dynamique de l'innovation web, ses jeunes ingénieurs inventifs et ses problématiques des réseaux ouverts. Les méthodes et les instruments de gestion ou de traitement de l'IST en France comme dans la plupart des pays avancés représentent une école de rigueur en appliquant, depuis les données de départ jusqu'aux produits

finalisés (essentiellement à l'époque des rapports circonstanciés via des commandes d'organismes publics ou privés), une série de contrôles qualitatifs pour vérifier la pertinence des contenus, veiller à l'exhaustivité des corpus identifiés et vérifier l'exactitude des calculs appliqués. Cette exigence de rigueur peut conduire d'ailleurs à retravailler régulièrement, et de façon manuelle, les données de départ (ce que l'on désigne par *curation des données*), en particulier lorsqu'il s'agit des données d'affiliation (normalisation de l'orthographe des noms propres et des acronymes, des données temporelles ou géographiques, recherche des doublons, qualification des données manquantes...). Une minutie peut-être héritée de l'art du catalogage en bibliothèque... et à laquelle j'eus quelques difficultés à me plier.

Cependant, au détour des années 2000, l'IST commençait à être impactée par l'avènement du Web et de ses applications distribuées, notamment autour des enjeux de l'interopérabilité des bases entre elles via de nouveaux procédés de connexions (les techniques d'API et de *data fusion* en particulier). L'aspect qui m'intriguait le plus n'était pas celui de la « mise en connexion » des gigantesques silos d'informations scientifiques qualifiées avec la dynamique du Web (il était grand temps de multiplier les initiatives en Europe !), mais le potentiel d'exploration que représentaient les bases des éditeurs avec leurs batteries de métadonnées qui me donnaient le vertige. Une base propriétaire comme Scopus, un *leader* mondial de l'IST, indexait chaque année plus d'un million de nouveaux articles et de documents scientifiques à travers un modèle de données comportant plus de 220 rubriques ! Pour le dire autrement, j'abandonnais quelque temps le Web et ses masses dynamiques (au point de faire de sa forme un objet d'étude) pour me plonger dans un univers a priori plus restreint en volume, mais beaucoup plus riche de différentes dimensions dont les corrélations me semblaient si nombreuses... Ainsi, la première réunion de l'équipe « véhicule autonome » a consisté à définir lesquelles nous pourrions rapidement exploiter et le plus efficacement possible dans le modèle de données de Scopus (mais le WoS de Thomson Reuters n'aurait pas modifié la méthodologie adoptée).

À cette richesse des données de départ, le domaine de l'IST intègre aussi une pratique historique de la cartographie relationnelle sous la forme de *scientogrammes*, l'une des traditions de l'IST. Je me disais donc qu'il y avait là un filon à exploiter, une « niche » où se concentraient quelques pistes inédites pour moi et utiles à notre projet. Historiquement, en effet, on doit les premières cartes à base de graphes à la *scientométrie*, le domaine de la « mesure de la science » fondée massivement sur l'analyse de la façon dont les articles scientifiques (et donc, par agrégation, les revues ou les champs disciplinaires) *se citent* les uns les autres, tout comme pour le Web dont le lien hypertexte est la clef de voûte. Les productions scientifiques au sens large (articles mais aussi revues, ouvrages, rapports, synthèses...) ont la particularité

d’être bâties sur un système riche et savamment distribué de références qui intègre ainsi un document dans un système plus vaste dont la scientométrie contribue à analyser les formes. Un corpus est donc constitué d’une série de notices bibliographiques renseignant sur de multiples aspects du document :

Image 15.1

Indicateurs de contenu / mots-clés	Auteur(s)
Titre	Affiliation(s) (laboratoires, équipes, projets...)
Résumé	Date(s)
Table d'indexation (full-text)	Type(s) de document (article, communication...)
Free-tagging (mots-clés d'auteurs)	Citations (entrantes/sortantes)
Descripteurs (vocabulaire contrôlé, termes d'index éditeur...)	Co-citations (co-citant/co-cité)
Catégories et classifications thématiques générales	Références bibliographiques et Couplage bibliographique
	Reuves (groupement d'articles)
	Flux de citations entre revues

Interface de Scopus.

Les citations directes d’auteur à auteur, les références bibliographiques, les revues communes ou les concepts partagés finissent par générer une géographie particulière de « la science », s’accumulant au cours du temps comme des couches sédimentées. C’est surtout l’étude des flux de citations entre documents scientifiques qui a concentré l’attention des chercheurs et des ingénieurs, notamment depuis les travaux liminaires en *information sciences* de Derek de Solla Price⁵ dans les années 1960. Le fait d’être cité dans un article constitue pour un scientifique une façon de mesurer son « rayonnement », son influence ou son autorité intellectuelle dans un domaine scientifique ou même à l’échelle de plusieurs disciplines. Mais la concentration des

5. D.J. de Solla Price, “Networks of Scientific Papers”, *Science*, 149 (3683), 1965, pp. 510-515.

liens (entrants) de citation permet aussi de dessiner à grande échelle les éléments structurants de « la science » avec ses auteurs comme ses concepts centraux⁶.

Image 15.2

Cartographie numérique de la recherche en chimie en France en 2009 à partir des données du WoS (Linkfluence, WebAtlas, 2010). Les chercheurs en chimie des laboratoires français ont publié en 2009 près de 6 000 articles, diffusés dans plus de 750 revues et sources d'information. Ces articles, et donc ces sources d'information en chimie (et dans d'autres disciplines comme la physique ou la biologie) sont connectées entre elles par les 23 722 liens d'intercitation que produisent les chercheurs en chimie des laboratoires français. Ces informations extraites du Web of Science donnent une idée assez précise de l'ampleur des publications en chimie, mais aussi des spécificités de la recherche française dans le domaine (notamment à travers les liens de citations que produisent les chercheurs français entre eux). Ces données du WoS ont été intégrées ici dans une technologie cartographique de pointe, dédiée normalement à des tâches de veille de sources web : interface interactive zoomable, distribution spatiale des sources, couleurs attribuées à chaque sous-domaine de la chimie (analytique, organique, matière...), sélection et visualisation pour chacune des 750 sources de publications du réseau de liens de citation, accès aux URL des revues, etc. Le passage des données du WoS à des technologies de veille élaborées est donc aujourd'hui possible et ouvre les « cartographies des sciences » aux fonctionnalités des interfaces numériques et des réseaux. En l'état, le dispositif constitue un espace efficient de synthèse de la recherche française en chimie et peut être potentiellement remis à jour chaque année (34).

6. K. Börner, "Mapping the Structure and Evolution of Science", dans *Symposium on Knowledge Discovery and Knowledge Management Tools at NIH Natcher Conference Center, Bethesda February 6, 2006*.

Les méthodes scientométriques me rappelèrent rapidement que l'ossature du Web est constituée des briques élémentaires que sont les liens hypertextes qui empruntent largement à l'esprit et à l'écriture des citations de textes à textes ou des citations bibliographiques. Le Web, comme les bases documentaires dédiées à la science, sont tous les deux nourris du projet commun de connaissances scientifiques ouvertes et partagées. Dans les deux cas, la *façon de se saisir* d'un corpus relève de la même méthode. En IST, l'étude de la distribution des citations entre documents passe par la production de structures de graphes à partir desquels on peut calculer des propriétés du même type, voire similaires, à celles déployées en *web mining* : *ranking* selon différents critères (comme le nombre de liens de citation ou toutes les formules associées), recherche de *clusters* ou de communautés, différenciation de dates dans les données pour appréhender la science comme une série de flux historiques, croisement des citations et des descripteurs de contenus, recherche de réseaux de coopération entre établissements ou organismes de recherche, etc. Au-delà des calculs, les structures de graphes peuvent être aussi visualisées sous forme de *scientogrammes* (on n'utilise guère le terme de « carte », jugé trop polysémique dans la tradition de l'IST française), un terrain où excelle Gephi avec ses algorithmes de spatialisation.

Le terrain de jeu semble donc en partie le même, des méthodes de l'IST à l'exploration du Web. Mais si les domaines sont connexes, les objectifs diffèrent sensiblement. Autant le Web est un espace ouvert et dynamique dont on ne connaît pas les limites, ou même l'envergure, autant l'IST est gouvernée par les questions d'exhaustivité, de clôture et de pertinence des corpus⁷. Mais la différence essentielle tient selon moi aux objectifs des méthodes d'analyse qui ont été historiquement assignés à l'IST : l'analyse des produits officiels et formalisés de la science à partir de mesures statistiques et l'observation de la structure de la science conçue comme une activité organisée et validée de production de connaissances. La notion de *structure* est à prendre en scientométrie au sens large de « géographie de l'information scientifique » et peut concerner tout aussi bien les frontières de disciplines que la distribution des concepts et des références bibliographiques, ou encore la métrique des procédés cartographiques, les données sur lesquelles elle s'appuie, les scénarios locaux ou globaux d'évolution (*dynamics, longitudinal mapping*), la détection de champs de connaissance en émergence (*research fronts*) ou encore la question de

7. Trois concepts qui posent problème dans la perspective du *web mining*. Voir chapitre 4 du présent ouvrage, « Agrégats ».

la nature de la recherche « pluridisciplinaire ». Dans cette perspective, on comprend que les scientogrammes alimentent une longue tradition spéculative sur la forme du savoir et des connaissances, occupant ainsi une place de choix dans la tradition américaine en philosophie des sciences ou en épistémologie.

Observation de « la science », mais aussi mesures et évaluation des acteurs et des organisations (laboratoires ou établissements), construction d'indicateurs de performance, appui logistique aux méthodes de « management de la recherche », aide à la décision en termes d'orientation des budgets. Alors que les « cartes », les graphes ou leurs techniques de visualisation n'occupent qu'une place périphérique en *web sciences*, on peut aisément affirmer que rares sont les domaines de recherche comme la scientométrie qui ont autant interrogé la nature, les formes ou les fonctions des graphes comme mode de navigation et d'évaluation (*mapping sciences*) dans les connaissances scientifiques formalisées. À l'époque où je découvrais la rigueur et la discipline méthodologique de ce domaine, quelques projets aboutis de cartographie de l'IST émergeaient tels Map of Science⁸, SCImago⁹ ou Eingenvector¹⁰ et ont contribué à une forme de renouveau technologique dans le domaine de la cartographie comme instrument de traçabilité de l'activité scientifique.

Dans cette « fièvre évaluative » des années 2000, la cartographie de l'information scientifique joue un rôle central dans le déploiement de projets et d'initiatives à l'échelle mondiale qui comprenaient aussi bien le fameux facteur d'impact¹¹ que l'évaluation des orientations stratégiques nationales et des financements, la gestion coordonnée des connaissances (par exemple, leur concentration en pôles), la valorisation de la production intellectuelle sous forme de capital ou de brevets, les comparaisons internationales, la définition de tendances lourdes en matière de R&D, la mise en valeur des partenariats public-privé... Sommé de participer activement à un renouveau industriel nécessaire, l'univers de la recherche s'est donc vu cartographié sous une multitude d'angles à des fins de stratégies des politiques publiques de soutien ou d'orientation.

Dans cette perspective, la capacité de synthèse de l'instrument cartographique a permis d'isoler des propriétés macroscopiques de connaissances scientifiques dont

8. <http://mapofscience.com/>.

9. <http://www.scimagojr.com/>.

10. <http://www.eigenfactor.org/>.

11. Le facteur d'impact est un calcul appliqué aux revues scientifiques et constitue un indicateur de rayonnement. Il est basé sur le nombre moyen de citations de chaque article publié dans cette revue (citations internes et externes à la revue). Le facteur d'impact a été conçu par Eugène Garfield, le fondateur de l'Institute for Scientific Information (ISI) maintenant propriété de Thomson Reuters. Chez l'éditeur, l'ensemble des données de citation sont publiées chaque année dans le *Journal Citation Report*.

la production augmente sans cesse : avec plus d'un million de publications recensées chaque année par les grands éditeurs (comme Elsevier et Thomson Reuters), on comprend combien les cartes de la science constituent des outils puissants pour interroger les propriétés sociales, éditoriales et stratégiques de l'activité scientifique à différentes échelles, des « cartes mondiales » aux micro-visualisations dynamiques générées par auteur ou par laboratoire. Et les scientogrammes accompagnent aussi l'exploitation des informations archivées dans des bases comme le Web of Science et Scopus), préparant souvent la mise au point de nouveaux services payants chez les éditeurs. Ici, comme ailleurs, les États-Unis jouent un rôle moteur dans ce renouveau pour la cartographie des sciences : dès 1987, un rapport de la National Science Foundation préconisait un financement immédiat et à long terme du domaine de la visualisation scientifique et des sciences. En 2003, l'US National Academy of Sciences a organisé une conférence spécifique et fondatrice dédiée au domaine de la *mapping science* (voir l'excellente introduction au domaine de Katy Börner *et al.*¹²). Un dispositif de communication comme Places and Spaces¹³ vient aujourd'hui confirmer au niveau mondial l'intérêt croissant porté aux cartes des sciences, voie ouverte dès 1981 par la publication du premier Atlas scientifique par l'Institut for Scientific Information (ISI) ou par la lignée des travaux scientographiques depuis Price (1965) en passant par Small et Sweeney (1985) jusqu'aux travaux actuels de Loet Leydesdorff ou Richard Klavans¹⁴.

Au moment où démarrait notre projet de cartographie du domaine des recherches autour du *véhicule autonome*, ces cartes de la science présentaient un intérêt méthodologique (si l'on pense à la rigueur apportée à la constitution des corpus en IST), mais je n'étais pas convaincu de les produire sur la base des citations. Nous avions d'abord besoin de comprendre le thème et ses nombreux sous-ensembles probables, de produire une carte synthétique des concepts eux-mêmes ou des thèmes de recherche, permettant de décrire un domaine dynamique de recherche comme une sorte de puzzle composé de multiples éléments que l'on pourrait mettre bout à bout. En somme, une boussole conceptuelle, et la richesse des métadonnées des notices Scopus le permettaient, notamment à partir d'une batterie de descripteurs

12. K. Börner, C. Chen et K.W. Boyack, "Visualizing Knowledge Domains", *Annual Review of Information Science & Technology*, 37 (1), 2005, pp. 179-255.

13. <http://www.scimaps.org/>.

14. R. Klavans et K.W. Boyack, "Toward a Consensus Map of Science", *Journal of the American Society for Information Science and Technology*, 60 (3), 2009, pp. 455-476, <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20991>.

contrôlés manuellement : les *index terms*. Deux interrogations me poussaient ainsi à mobiliser des *concepts* (associés à des acteurs, des dates, des organismes) plutôt que les *citations*. L'étude des distributions de liens de citation me semblait plutôt nous orienter vers la cartographie d'un champ social particulier, celui des auteurs et de la façon dont ils en construisent les références (ou les autorités). Il m'apparaissait que, quand un auteur en cite un autre, il paie rétrospectivement une sorte de tribut à ses prédécesseurs ou à ses contemporains en les reconnaissant. L'accumulation de tous ces liens permet effectivement de construire rétrospectivement la science comme une série de concentrations (les objets de recherche en cours) ou de flux historiques (les évolutions successives des domaines scientifiques). C'est tout le principe de l'utilisation du Science Citation Index (SCI) vendu par Thomson Reuters, et que je préférerais appeler le « livre des citations ». Ce « catalogue » concentre les liens de citation de revues à revues (en agrégeant les liens de citation d'auteurs à auteurs) et permet donc de les classer (avec sa fameuse classe « A »), mais aussi de décomposer les productions scientifiques en disciplines et en sous-disciplines. Si le projet de cartographie globale de « la science » a longtemps fasciné les scientométriciens, un autre point m'interrogeait : les graphes à base de citation supposent que les liens sont *orientés*. Un chercheur A cite un chercheur B dans un article, mais souvent B ne cite pas A dans le sien, principe à partir duquel on peut extraire pour chaque article les liens entrants (*cited document*) et les liens sortants (*citing document*). Que deux documents, ou plus, en citent un troisième et l'on fait émerger par l'analyse des co-citations des hiérarchies « verticales » qui indiquent clairement le cœur de discipline ou, pour des champs de recherche plus restreints, des problématiques particulières.

Nous avions plutôt besoin, dans un premier temps, d'une coupe longitudinale du domaine, fondée sur une série d'expressions représentatives, quitte à réintroduire plus tard les métadonnées associées aux auteurs. Ces derniers, d'ailleurs, nous apparaissaient plutôt comme des *acteurs* situés *dans le temps et dans l'espace* (leurs lieux de travail et leurs laboratoires, métadonnées présentes dans le corpus que nous allions constituer) que l'on pourrait tous cartographier, et même associer à des sources web comme les réseaux sociaux. Notre corpus comportait au final 481 notices bibliographiques consacrées à la recherche française dans le domaine du véhicule autonome¹⁵. Le premier temps de notre méthodologie était constitué par la phase d'extraction des notices. Il fallait pour cela interroger la base Scopus à partir d'un jeu de termes enrichi par phases successives d'expansion, jusqu'à l'obtention d'un jeu de données suffisamment « exhaustif ». À partir d'une première expression (ou d'une série d'expressions) relative à notre thème (*autonomous vehicle*) et choisie par notre groupe de travail pour son intérêt supposé (nous sommes accompagnés de deux experts du domaine), le processus consiste à intégrer petit à petit (par

15. Du moins, des articles écrits par des chercheurs membres d'un laboratoire français.

degré de « profondeur ») d'autres expressions avec lesquelles l'expression « véhicule autonome » peut être associée. On pense, spontanément, à bien des thématiques comme « robotique », « intelligence artificielle », « autonomie » ou « décision »¹⁶. En associant chacune de ces expressions à *autonomous vehicle*, nous avons fait varier les requêtes sur Scopus (sur les dimensions « *keywords* », « *abstracts* », « *article title* ») pour obtenir au final 740 notices avec un taux élevé de redondance pour chacune des entrées. Par exemple, certaines notices bibliographiques sont présentes dans presque tous les résultats du requêtage. Après nettoyage des doublons et de quelques lignes mal renseignées par l'éditeur, nous avons construit notre travail sur 481 notices bibliographiques extraites de Scopus dans lesquelles au moins un auteur est associé à un laboratoire français.

Restait à construire le paysage sémantique particulier, et peut-être complexe, autour des recherches sur le véhicule autonome. Chacune de ces notices contient aussi des descripteurs de contenu (*index keywords*) décrivant l'apport scientifique général de l'article. Une fois le graphe de descripteurs produit (un descripteur est lié à un autre descripteur quand ils apparaissent ensemble pour qualifier le contenu d'un même article), nous avons sélectionné un jeu de 1 300 termes, expressions ou descripteurs de contenus reliés entre eux dans une structure de graphe à partir de leurs liens pondérés de co-occurrence. La première vue obtenue en utilisant l'algorithme Open-Ord laissait apercevoir pour la première fois l'organisation globale de connaissances scientifiques produites par les laboratoires français sur le « véhicule autonome » :

16. • *Automobile* • *Collision Avoidance* • *Communication* • *Computer Science* • *Control* • *Fleet* • *Highway* • *Human Machine Interaction (HMI)* • *Image* • *Localisation* • *Multi-Agent System* • *Navigation* • *Obstacles* • *Optimization* • *Optimization Realtime* • *Real-time* • *Robotics* • *Simulation* • *Software* • *Software Trajectory* • *Trajectory* • *Transportation* • *Vehicle dynamics*

Image 15.3

(a) spatialisation avec OpenOrd (Gephi) des 13 000 *index terms* et de leurs relations de co-occurrence dans les 489 notices du corpus retenu ; (b) regroupement des 27 653 liens de co-occurrence en fonction de 7 *clusters* produits avec Open Ord / filtrage des 1 300 termes ayant une occurrence supérieure à 25, soit 689 termes conservés ; (c) regroupement des nœuds en 7 *clusters* et attribution des couleurs.

Le processus d'analyse des données faisait clairement apparaître un domaine « central » (*control theory*) et six domaines « périphériques » (*computer systems*, *oceans and marine science*, *propulsion/flight*, *entries and aerodynamics*, *space autonomy*, *control systems*). À l'évidence, un cœur dense de recherche concernait la « théorie du contrôle », donc l'intelligence artificielle et la robotique en grande partie, tandis que les 6 satellites concernaient plus précisément différents champs d'application, dont trois le domaine du ciel et de l'espace (missiles, drones, aviation, missions spatiales), un l'océan (sous-marins, robots automatisés d'exploration) et les deux autres les systèmes informatiques associés (disons, leur architecture et leurs propriétés) ainsi que la dimension communication et réseaux auxquels ils seront connectés. Tout cela constituerait l'ossature de la boussole thématique que nous attendions dont ces 7 *clusters* permettraient de s'orienter en première approche dans les connaissances scientifiques.

Image 15.4

Étude sur le véhicule autonome (détail).

Au-delà de ce premier niveau géographique, le domaine central méritait d’être exploré pour lui-même, indépendamment de ses satellites (qui, par ailleurs, rattachent notre objet d’étude à bien d’autres domaines connexes de la recherche dans la carte générale des savoirs). L’examen attentif de ce noyau conceptuel constitué de « *control theory and applications* », repose sur l’utilisation de l’algorithme de spatialisation Force Atlas 2 et une distribution des couleurs en fonction des frontières de *clusters* détectés par l’algorithme de détection de « communautés » Modularity avec Gephi. La taille des nœuds était réglée selon le degré de co-occurrence des termes dont les plus importants ont été affichés au premier plan, les autres sous forme de cartouches latéraux.

Image 15.5

Étude sur le véhicule autonome (détail).

En gris, figurent au centre les termes génériques du domaine, reliés à toutes les autres parties du graphe dont nous avons retiré les liens pour assurer une pleine lisibilité des zones déterminées. En jaune, au centre, le domaine global de *robotics/robots/algorithms* où figurent les modèles mathématiques et les algorithmes, des probabilités à la logique floue, les modèles non linéaires liés aux capacités de contrôle (mouvement, déplacement, opérations) de la décision en situation complexe et multivariées. Très lié à ce cœur, presque fusionné, apparaît en rouge la principale interface opérationnelle des algorithmes et du calcul : les formes de contrôles et les systèmes associés, déployés autour des domaines du déplacement/mouvement et du processus de décision en situation. Viennent ensuite deux domaines en vert à droite, les robots mobiles avec les questions des programmes de déplacement et de contrôle des trajectoires, et la navigation en milieu urbain avec ses propriétés attendues (route, rues, circulation...) – le domaine émergent des *cybercars* dont on parle tant aujourd'hui. Enfin, deux domaines en bleu à gauche : en haut, les questions de géolocalisation dynamique, des environnements augmentés/hybridés et en bas à gauche les familles de capteurs, achevant une boucle puisque les informations qui en sont issues sont traitées par... des algorithmes et alimentent des modèles mathématiques du transport autonome que l'on trouve concentré en jaune, au centre. La boussole nous semblait ainsi opérationnelle, nous permettant de synthétiser tout à la

fois les zones constitutives d'un cœur de recherche et, à un niveau plus général, les différentes branches représentant les grands domaines d'application : ciel, espace, océans, propulsion, aérodynamique.

La construction de la boussole thématique nous avait orientés vers la cartographie de l'activité scientifique elle-même, avec ses concepts, mais aussi ses acteurs et ses structures. À partir des notices bibliographiques, nous avons laissé derrière nous l'exploitation des liens de citation, les mesures, les classements, les filiations intellectuelles et la recherche des figures académiques d'un domaine très dynamique et en plein développement. Mais nous devions aller plus loin dans nos clichés cartographiques : tout nous incitait à croiser les différentes dimensions du corpus comme les auteurs, les sources et/ou les données d'affiliation aux structures de recherche (laboratoires ou établissements) via les adresses postales ou les noms des villes, pour essayer de rendre compte de la géographie sociale du domaine. Où les chercheurs conduisent-ils leurs expérimentations ? Qui sont les établissements qui contribuent à la recherche sur le véhicule autonome ? Où sont-ils situés en France, de façon concentrée en certains lieux ou bien distribués à l'échelle du territoire national ? Les notices de notre corpus comprenant les noms des auteurs des publications scientifiques, leurs thèmes mais aussi le nom et le plus souvent l'adresse postale de leurs laboratoires ou établissements. Nous avons transformé les codes postaux en coordonnées GPS. Le procédé rend ainsi possible la projection de l'ensemble des autres métadonnées sous forme de points sur une carte géographique quand il s'agit de structures physiques comme des laboratoires, mais aussi de liens d'association pour signifier des coopérations ou le partage de concepts. Évidemment, étant donné notre domaine large du « véhicule autonome », il n'y eut aucune surprise à contempler la carte de France avec les unités ainsi identifiées. Les différents aspects des recherches autour du véhicule autonome couvrent l'ensemble des grands pôles nationaux de recherche avec la présence des universités, mais aussi d'industriels comme Hitachi Europe, Renault (technocentre), Véhipôle en Bretagne, Siemens, Thomson et les divisions de Thalès Airborne Systems – Underwater Systems – Alenia Space... On retrouve, notamment, une tripartition qu'annonçait notre boussole conceptuelle : mers et océan avec des concentrations d'acteurs en Bretagne et en PACA (École navale, Alstom Transport, Bassin d'essais des carènes, DCNS, la DGA et le Groupe d'études sous-marines de l'Atlantique – GESMA – géoazur, un observatoire océanologique, les centres IFREMER, l'Institut de physique du globe de Paris, le laboratoire de

géosciences marines...), espace (Aérospatiale, Agence spatiale européenne, Astrium Satellites, le CNES) et l'aérien (SupAero, Dassault Aviation, EADS...).

Image 15.6

Étude sur le véhicule autonome (détail).

Nous avons sous les yeux la colonne vertébrale de la recherche militaire et industrielle, privée et publique, dans le domaine du « véhicule autonome », le champ incluant aussi bien les missiles que les sous-marins, les avions ou la R&D du secteur automobile ! Les données de Scopus illustrent donc certaines propriétés connues de l'espace industriel et économique de la France, en particulier la concentration des acteurs en région Île-de-France. Nous avons ainsi produit une vue particulière sur la région, laissant apercevoir la concentration de grandes structures de recherche à Paris (Mines ParisTech, l'École des ponts ParisTech, l'École polytechnique, l'ENS de Cachan, l'université Pierre et Marie Curie et ses structures de recherche...), mais aussi dans la partie ouest/sud-ouest (le fameux « plateau de Saclay »).

Image 15-7

Étude sur le véhicule autonome (détail).

Le géoréférencement des métadonnées ouvrirait de nombreuses perspectives, presque impossibles à toutes identifier au moment de notre projet et toutes porteuses de pistes nouvelles pour l'époque. Le nombre et la qualité des données associées à nos 481 notices bibliographiques permirent d'apercevoir l'énorme potentiel de croisement des différentes dimensions de l'information en IST. Ainsi, nous avons ébauché le projet de cartographier la zone sud-ouest de Paris autour de Saclay pour mesurer l'étendue des coopérations de recherche via la publication d'articles communs à des chercheurs provenant des structures présentes dans cette zone géographique. Le géoréférencement permettant de constituer un sous-corpus de laboratoires et de chercheurs implantés dans notre focus géographique, il devient alors possible de produire un graphe où les chercheurs (regroupés en laboratoires ou en établissement) peuvent être reliés par les

index terms qui nous avaient permis de construire la boussole thématique. Il devenait ainsi possible de comprendre sur quelles thématiques de recherche coopéraient les établissements autour du plateau de Saclay.

Image 15.8

Étude sur le véhicule autonome, visualisation exploratoire.

Depuis les méthodes traditionnelles de l'IST et l'obsession des citations, nous glissons petit à petit vers des méthodes et des dispositifs du domaine en pleine expansion des cartographies d'informations en intelligence territoriale. À Saclay, comme ailleurs, il existait bien un réseau local de production des connaissances scientifiques dont on pouvait enfin observer la dynamique. Évidemment, l'introduction de la temporalité ou celle des informations issues des appels à projets (AAP) des agences nationales, européennes ou régionales auraient encore enrichi la perspective comme je l'avais expérimenté avec le projet VIZIR¹⁷. Mais se dessinaient aussi d'autres pistes qui nous menaient bien au-delà du croisement des métadonnées entre elles : avec les informations qu'elles nous livraient (noms d'auteurs et de structures, lieux, années, mots-clés...), il était possible de connecter notre corpus à bien des services web. Nous avons ainsi projeté directement sur Google Map nos fichiers qui contenaient des données GPS, développé un petit programme pour aller chercher sur LinkedIn les profils professionnels des chercheurs, indexé les contenus des sites web des laboratoires ou, encore, comparé différents laboratoires sur la base de leurs recherches autour du *véhicule autonome*.

17. Voir le chapitre 16, « Les écosystèmes d'innovation ».

Image 15.9

Étude sur le véhicule autonome, projection géographique des principaux acteurs.

* * * * *

Nous n'étions pas les seuls à explorer les pistes qui permettaient techniquement de connecter plusieurs sources d'information externes aux données qualifiées de l'univers de l'information scientifique. La maturité technique du Web permettait d'expliquer cette fièvre de la connexion entre différents silos d'information, le développement des API en particulier qui permettait d'interroger des BDD à distance. Certains grands acteurs de l'IST comme la base Medline (US National Library of Medicine¹⁸) commençaient par ailleurs à fournir des accès gratuits à des masses de données et à une bonne partie des publications majeures en sciences du vivant. Mais le plus remarquable dans ces évolutions rapides de l'IST fut le développement des réseaux scientifiques sur le Web, les *réseaux scientifiques 2.0*. Bien des acteurs avaient franchi le pas en proposant en ligne des systèmes de recherche et de partage de documents scientifiques comme ResearchGate¹⁹, Sciencefeed²⁰, MyScienceWork²¹.

Mais cette rencontre, ou plutôt cette hybridation des données de l'IST et des technologies web me semblait surtout prometteuse pour interroger, comme l'avait fait la scientométrie, certains des invariants d'une géographie globale de la science, comme une carte générale des savoirs avec ses instruments de navigation dans

18. <https://www.nlm.nih.gov/bsd/pmresources.html>.

19. <http://www.researchgate.net/>.

20. <https://twitter.com/sciencefeed>.

21. <https://www.mysciencework.com/>.

les données. À défaut de bénéficier d'une infrastructure de calcul et de données à l'échelle mondiale, je me suis tourné vers l'univers qui m'était le plus proche, celui de mon université de technologie²². Je partais dans cette expérience avec l'idée de produire le même type de paysage intellectuel que celui de la recherche sur le véhicule autonome. Nous avons donc suivi une méthodologie identique et en extrayant là aussi les données depuis Scopus. Nous sommes rapidement arrivés à la constitution d'un corpus de 5 200 notices bibliographiques renvoyant aux 5 200 articles scientifiques répertoriés dans la base et publiés depuis 1972 par des chercheurs issus des laboratoires de l'UTC, l'année de naissance de l'université. Ces notices contenaient 15 000 termes d'index que nous avons filtrés de façon à n'en conserver que 1 000 pour notre carte (en ne retenant que ceux qui apparaissent au moins 7 fois dans le corpus). Une structure de graphe comprenant 1 000 termes a été générée où ne figuraient que les termes d'index reliés entre eux par co-occurrence entre chaque notice (un terme A est lié à un terme B s'ils apparaissent tous les deux comme descripteurs d'une même notice). La structure de graphe a ensuite été spatialisée et décomposée en 81 *clusters* (Modularity) qui représentent les grands thèmes de recherche des laboratoires de l'UTC. Les 1 000 termes d'index de Scopus qui servent à décrire le contenu des 5 200 publications depuis 1972 ont donc été agrégés en 81 thèmes qui concentrent chacun un vocabulaire spécifique relativement aux autres. Les labels de nos 81 classes sont constitués par le terme ayant le meilleur score de co-occurrence de la classe, hormis pour les *clusters* en vert dans la carte où nous avons fait appel à un chimiste pour les valider et, parfois, les modifier pour les rendre plus pertinents.

Image 15.10

Réseau de collaborations scientifiques à l'UTC (détail).

22. <https://ateliercartographie.wordpress.com/2014/03/21/cartographie-des-publications-de-lu-t-c-depuis-1972-premiere-version/>.

La vue générale, accompagnée de ses vues secondaires, me semble décrire ce type de géographie des connaissances qu'annonçait l'étude sur le *véhicule autonome*.

Image 15.11

Réseau de collaborations scientifiques à l'UTC.

La distribution des couleurs respecte dans ses grands traits les compétences scientifiques de nos laboratoires, avec des « frontières » qui ont fait l'objet de nombreux débats dans notre groupe. Ces débats ne sont pas anodins et l'on peut y voir l'une des caractéristiques de la recherche technologique en école d'ingénieurs : au-delà des disciplines « verticalisées » par tous les effets de la classification (par exemple, un chercheur est obligatoirement rattaché à une discipline de référence d'un point de vue administratif), les thèmes de recherche et d'expérimentation reposent souvent sur le croisement de compétences différentes. Il se peut, d'ailleurs, que la fameuse « pluridisciplinarité » tant recherchée dans la recherche française représente la clef de voûte de la recherche technologique : travailler, comme à l'UTC, sur le « foie artificiel » ou les « flottes de drones » supposent que soient associés, d'une façon ou d'une autre, la biologie, l'ingénierie cellulaire ou les procédés d'un côté ; de l'autre, l'informatique, les matériaux et les spécialistes de l'énergie embarquée. Plus que tout, le poster synthétique que nous avons imprimé me semblait annoncer le type de *dashboard* numérique que j'ai toujours voulu voir installé, dans un hall d'entrée ou un bureau de direction.

Image 15.12

On pourrait y recenser et y positionner aussi bien les thématiques de recherche que les laboratoires, y rassembler les connaissances produites et les interroger via un moteur interne de recherche sur les publications elles-mêmes ou donnant accès aux rapports d'orientation stratégiques ou d'évaluation comme ceux produits pour notre Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (HCERES). Une frise chronologique permettrait d'identifier dynamiquement les thèmes de recherche émergents, ainsi que des informations sur les coopérations régionales, nationales ou internationales. Et l'on imagine rapidement combien serait également utile une carte géoréférencée des acteurs et des organismes, le tableau de veille intégrant aussi bien la cartographie des contenus que la dynamique sociale et territoriale qui l'accompagne. En hybridant ainsi données qualifiées de l'IST, données et services externes issus du Web et les potentialités offertes désormais par le numérique, on pouvait enfin traiter de façon fine et exhaustive la traçabilité de l'activité scientifique de la bibliothèque et ses ouvrages que l'univers foisonnant du Web et de ses innovations. Les pistes d'hybridation que nous avons explorées nous permirent d'entrevoir les mille et une façons dont les connexions pouvaient être établies entre l'univers ordonné de la bibliothèque et des revues, où se concentre lentement le travail patient de publication scientifique, et l'univers d'un réseau foisonnant et dynamique qui brasse les contenus, les auteurs, les dates ou les lieux de façon continue. La dernière fois que se réunit notre petite équipe, près de la bibliothèque, s'imposait l'impérieuse nécessité de poursuivre par ailleurs l'exploration de la vaste architecture distribuée du Web qui ne pouvait pas relever dans mon esprit du hasard ou du chaos, mais d'un agencement ou d'une sorte de plasticité structurale qui le rendait techniquement partout présent, et si puissant, étendant sa toile petit à petit jusque dans les lieux que l'histoire avait jusqu'ici préservés.

Chapitre 16. Les écosystèmes d'innovation

« Biodiversity research typically focuses on species richness and has often neglected interactions, either by assuming that such interactions are homogeneously distributed or by addressing only the interactions between a pair of species or a few species at a time. In contrast, *a network approach provides a powerful representation of the ecological interactions among species and highlights their global interdependence*. Understanding how the responses of pairwise interactions scale to entire assemblages remains one of the great challenges that must be met as society faces global ecosystem change », Jordi Bascompte, “Disentangling the Web of Life”, *Science*, 2009.

J'ai récemment assisté à une conférence sur la « mise en place d'un écosystème d'innovation ». Étant personnellement tourné vers le développement des *start-up* du numérique, ou plus généralement vers l'innovation technologique liée à la culture des *data*, c'est le type d'événement vers lequel je porte de plus en plus mon intérêt. Je collectionne aussi quelques articles parus dans la presse spécialisée ou les magazines, où l'on vante tous les bienfaits d'une organisation « en écosystème » pour « la compétitivité de nos entreprises », « la santé de notre économie » et le « dynamisme des territoires ». L'expert qui animait la présentation était issu d'un cabinet de conseil et d'audit spécialisé dans l'accompagnement des acteurs publics en matière « d'innovation » et de « valorisation économique des territoires ». Le début de l'exposé, accompagné de *slides* esthétiques, fut prometteur : l'expert présenta son travail d'identification des acteurs locaux de l'innovation (université, laboratoires de recherche, structures juridiques de valorisation comme les fondations, les différents types d'acteurs industriels – du plus petit au plus grand : *start-up*, petites entreprises, PME, grands groupes industriels et filiales) sans oublier les collectivités territoriales et leurs services. Il insista ensuite sur les mécanismes d'aide et de soutien de l'État mobilisables pour développer un *écosystème local d'innovation*, en soulignant les termes de « coopération », de « mobilisation des acteurs », de « diffusion de l'information », de « travail en réseaux », « d'émergence et de captation d'idées nouvelles ». Les perspectives ouvertes par les programmes nationaux tournés autour de la « compétitivité pour exporter », du renouveau de l'innovation et des mécanismes du « Grand Emprunt » et des « emplois de demain » furent maintes fois citées, l'auditoire semblant acquiescer. J'attendais donc ses propositions et ses conseils d'accompagnement ; en gros, les recettes qui permettaient de créer de toutes pièces

un « écosystème technologique innovant », de suivre son développement et, peut-être, d'en évaluer les formes de plus-value.

C'est là que l'exposé devint passionnant : plus l'expert avançait, moins les graphiques projetés me semblaient adéquats au propos. Au fur et à mesure que le discours se déroulait, il peuplait ses graphiques de flèches *descendantes*, les acteurs de l'innovation étant distribués dans l'espace de façon *hiérarchique* – presque arborescente. Ils semblaient quasiment alignés. Quand il aborda les notions de « créativité collective », de « collaboration », de « création d'entreprises et de *start-up* », il les présenta comme le résultat de son alignement. Le *slide* suivant fut formidable : l'alignement fit place à une forme circulaire où figuraient les mêmes acteurs connus, une sorte de « ronde » qui semblait suffire à réduire le concept d'écosystème ou de système d'interactions vertueuses¹. Il insista alors sur le concept de « compétitivité » en pointant les acteurs de l'innovation et plaça en bas de son schéma général un grand trait fin censé symboliser le « développement économique territorial ». J'eus l'impression du décalage manifeste que produisaient les schémas de plus en plus rigides et le propos du discours, comme si l'expert poussait au maximum les possibilités de PowerPoint pour arriver à circonscrire dans son schéma des concepts qui semblaient lui échapper. L'ordonnancement des éléments et son système de flèches (typiques de ces instruments de communication que sont les *slides*) m'apparaissaient réduits pour transcrire toute la richesse et la complexité potentielle des mécanismes qui règlent un *écosystème*, qu'il soit naturel ou technologique, du moins tels que je les imagine. Cette impression désagréable vint peut-être du fait que je suis attentif, pour en avoir fait l'expérience sur le Web pendant des années, aux questions de « schéma » censé représenter son objet.

Mais il manquait manifestement à l'intervenant ce genre d'instrument que sont les graphes, ou même autre chose, pour rendre compte de concepts aussi riches et prometteurs que ceux attachés à la notion d'*écosystème naturel* : la *diversité* des *connexions* entre espèces, la question de l'*émergence* ou de *coévolution*, celle de la *résilience*, de *biotopes* ou de *territoires* qui ne peuvent pas être réduits symboliquement à une ligne en bas d'un schéma. À la fin de l'exposé, l'auditoire semblait pourtant satisfait et, en quelque sorte, *rassuré* par ce schéma qui me semblait pourtant confus et réducteur. Les acteurs essentiels du schéma étaient *connus* (en général des organismes publics et des grandes entreprises), leurs relations d'association ou de hiérarchie *conservées*, les termes *attendus* (« marché mondialisé », « compétitivité », « montée en gamme des produits et

1. Je conseille de taper sur Google « écosystème d'innovation », puis de sélectionner « images » pour se rendre compte de la fortune de la « ronde des acteurs » en matière d'innovation !

services », « effort d'innovation » dans « la recherche et l'industrie »). Après quelques questions d'usage sur l'avenir de cette « économie de la compétitivité mondialisée », la conférence prit fin, me laissant plus que perplexe sur ce que je venais de voir.

Cet épisode me rappelle combien de temps, de données et d'instruments il m'a fallu pour commencer à esquisser un schéma simplifié du Web que j'aime à qualifier *d'écosystème documentaire*. Depuis, dès que l'on annonce le concept « d'écosystème », je suis averti de l'amoncellement de difficultés théoriques ou graphiques qui vont s'abattre rapidement quand on se penche sérieusement sur la question. Ce n'était pas l'objectif de l'intervenant, ou alors il avait réduit volontairement la portée de certains des concepts mobilisés, à commencer par celui *d'écosystème*. Car invoquer le concept *d'écosystème*, c'est se trouver rapidement confronté à un certain nombre de nœuds problématiques difficiles à résoudre, surtout si l'on vient de l'univers de l'administration – territoriale ou universitaire – ou de la stratégie industrielle, des univers où l'on reste rivé aux schémas rigoureux, aux organisations hiérarchiques, aux filières industrielles qui s'associent remarquablement par complémentarité dans une arborescence. Tout cela pouvait très bien fonctionner, de façon naturelle, dans un monde dynamique, en expansion, tourné vers le « plein emploi » et des nations ambitieuses avec leurs grands programmes industriels ou d'équipement. L'époque, peut-être heureuse, où le temps lui-même pouvait être découpé en phases d'un processus que l'on synthétisait sous la forme d'un tableau d'avancement dans les bureaux des « *managers* » ou des « décideurs ». L'avenir était fixé sur des horizons lointains que l'organisation hiérarchisée de l'entreprise et de l'État venait cadrer. Oui, mais voilà : le monde actuel est dans un état de mutation accélérée (dont la « crise » est un symptôme) et d'une redoutable complexité à appréhender. L'une des solutions pour un pays industrialisé comme le nôtre a été de miser massivement, il y a quelques années, sur la *compétitivité par l'innovation* dans un marché *mondialisé*. Et le concept d'écosystème d'innovation a été mobilisé pour décrire les modalités locales ou opérationnelles de la mise en œuvre de ce programme ambitieux de « montée en gamme » des produits et des services. Les universités, les laboratoires de recherche et les chercheurs eux-mêmes sont appelés à se mobiliser autour de cet objectif générique de « création de richesses et d'emplois ». Quelle région, quelle métropole, quel territoire ne possèdent pas aujourd'hui en France son écosystème d'innovation, sa pépinière d'entreprises, son incubateur, sa cantine numérique, son quartier de la création, son centre d'innovation ou de transfert technologique, ses

laboratoires « d'excellence », ses services de développement économique orienté vers les nouvelles industries ? Et la conférence à laquelle j'avais assisté était directement issue de cette politique actuelle.

Je reconnais la difficulté de la tâche et les efforts à fournir pour identifier les mécanismes complexes qui régissent un *écosystème*. La difficulté s'accroît d'autant plus que l'on se place, comme notre orateur, dans une perspective *volontariste* de développement des *écosystèmes d'innovation*. Il est ainsi devenu courant d'invoquer la notion d'*écosystème* en matière d'innovation et de politique publique, mais je ne peux m'empêcher de penser que l'on se contente souvent (volontairement ?) d'une définition floue du concept, comme s'il s'agissait d'en circonscrire la portée théorique et en faire l'instrument quasi marketing de politiques locales de développement économique présentées comme « nouvelles ». Et je me demande, de façon peut-être trop suspicieuse, s'il ne s'agit pas là d'une manière habile de faire perdurer le rôle ancien des *mêmes* acteurs dans une organisation toujours aussi hiérarchisée. Pourtant le principe de l'écosystème appliqué aux processus d'innovation a quelque chose de fascinant dans la remise en cause des organisations complexes hiérarchisées (l'université et ses laboratoires, les collectivités territoriales, les grands groupes industriels), de leurs méthodes de management, des rôles dévolus pour chacun ou des objectifs poursuivis par tous. Pour être franc, si l'on met réellement en application quelques-uns des enseignements majeurs que nous ont livrés les biologistes et les écologues, le concept d'écosystème appliqué à l'innovation a quelque chose de *subversif* pour les organisations qui nous gouvernent et les acteurs qui en sont responsables.

Ici, je ne peux masquer le *hiatus* ou la *contradiction* qui m'habite depuis quelques années entre les discours disons « officiels » sur l'innovation et les *écosystèmes* (et la réalité des premiers projets mis en œuvre, dont certains auxquels j'ai participé) et ce qu'évoque, par ailleurs, le concept « d'écosystème » avec son cortège de processus associés si l'on s'aventure à comparer *innovation technologique* et *biodiversité naturelle*. Je parie donc sur le fait qu'une *observation minutieuse* et une *cartographie méthodique*, à différentes échelles organisationnelles ou territoriales, des acteurs et des processus engagés dans les *écosystèmes d'innovation* (ou les projets présentés comme tels) permettront dans les années à venir de *manifester visuellement et de résoudre cette contradiction*. Selon les cas observés, a-t-on oui ou non affaire à un écosystème ? Doté d'un côté d'un appareillage descriptif (cartographique) et, de l'autre, d'une batterie de concepts ou de propriétés supposées d'un écosystème, il deviendrait ainsi possible d'en éprouver la réalité concrètement sur des territoires ou des processus. Au-delà des sentiments contradictoires qui m'animent personnellement, la cartographie des processus d'innovation et des territoires innovants représente un

domaine à peine émergent aujourd'hui, mais appelé à jouer un rôle majeur dans les années à venir. Les défis sont nombreux et divers, comme cela l'a été pour le Web ou d'autres univers ouverts et distribués : sur le plan technique, avec quels instruments ? Sur le plan méthodologique, avec quel(s) type(s) de données ? Sur le plan stratégique, s'agit-il seulement de description ou bien alors d'évaluation ? Évidemment, étant donné les interrogations qui m'animent aujourd'hui, j'aurais tendance à confondre les deux dernières dimensions, chacun ayant compris dans cet ouvrage tous les enjeux (pas seulement scientifiques) de la cartographie d'informations.

Mon intérêt pour le concept d'écosystème est relativement récent, mais il me permet d'articuler plusieurs pans de mon activité qui s'assemblent désormais comme un puzzle autour de cette notion problématique et maintenant centrale dans mes activités. Il y a tout d'abord ce genre d'expérience empirique produite par les projets d'innovation qu'il a fallu parfois conduire malgré un environnement hiérarchique qui ne laisse que peu de place aux initiatives spontanées, de surcroît dans un environnement caractérisé par la faiblesse des interactions entre une série d'acteurs censés coopérer pour accompagner les projets. Depuis la naissance de Linkfluence jusqu'à l'Atelier Iceberg en passant par Gephi, que d'aventures dans les méandres de l'innovation administrée ! Selon mon expérience, la mise en œuvre d'un écosystème d'innovation devrait commencer par un effort d'évaluation systématique des moyens mis en œuvre au regard des missions qu'ils sont censés permettre d'accomplir. Et la cartographie peut y contribuer, surtout dans le contexte où nos technologies permettent aujourd'hui de fusionner des silos informationnels différents (publications scientifiques, brevets, bases de connaissance, informations légales sur les entreprises, programme et projets des laboratoires de recherche...), de traiter des flux et des connexions sous forme de graphes relationnels, de géoréférencer des entités (notamment avec Google Maps) ou de visualiser dans une interface des évolutions temporelles. Au-delà des discours, et des promesses, un « écosystème d'innovation » se décrit et s'analyse en détail selon une série de mesures objectives et des concepts qui les justifient.

Pour préparer le terrain de ce vaste chantier de cartographie, et pour tenter de résoudre la contradiction qui m'habite entre ce que je vois se déployer et ce que je pense être *fondamentalement* un écosystème d'innovation, j'ai aussi été amené à consulter de nombreux documents de référence qui ont contribué à redessiner ces dernières années une politique industrielle construite sur l'innovation (rapport de la commission « Innovation 2030 » présidée par Anne Lauvergeon, « Refonder l'université, dynamiser la recherche. Mieux coopérer pour réussir » de Jean-Yves Le Déaut, le « Pacte pour la compétitivité de l'industrie française » de Louis Gallois, commissaire général à l'Investissement, les données disponibles sur les *investissements d'avenir*,

l'agenda stratégique pour la recherche, le transfert et l'innovation France Europe 2020, diverses informations à propos de l'ANR et du rôle de la BPI...). N'étant pas économiste ou féru de sciences politiques, chacun de ces documents ne recèle a priori guère d'intérêt particulier pour moi. Cependant, considérés ensemble, ils forment un tableau général cohérent et dessinent les contours essentiels d'une politique industrielle en apparence ambitieuse pour notre pays. Surtout, ils sont autant d'occasions de lire nos responsables politiques sur ce qu'ils entendent par « écosystème ». Et j'ai retrouvé dans cette masse de documents les termes de l'intervenant qui animait la conférence, évidemment. Et il n'y avait là rien de rassurant.

* * * * *

Pour conduire à bien ce projet de cartographie des écosystèmes d'innovation, il faut pouvoir disposer de deux instruments : un *appareil réflexif* construit autour de concepts et un *dispositif d'observation* capable de recueillir des données, quel qu'en soit le type ou la nature. Commençons par l'appareil réflexif qui doit accompagner l'analyse des données à partir de concepts ou d'un éclairage théorique permettant de saisir les différentes dimensions d'un écosystème. Sur ce point, si l'on tente de comparer *écosystèmes naturels* et *écosystèmes d'innovation* (du moins tels que doivent les concevoir l'orateur de la conférence ou les documents de stratégie économique), il apparaît rapidement que les premiers sont plus riches ou plus complexes que le second. La notion d'*écosystème naturel* semble contenir plusieurs formes d'organisation que les écologues tentent actuellement de décrire et d'identifier². Mais, évidemment, il faudrait être écologue soi-même pour percevoir toutes les nuances de ce qu'est un *écosystème* en tant que tel. Cependant, mon passage à la Fondation pour la recherche sur la biodiversité m'avait permis de m'intéresser aux différents aspects du principe au cœur duquel réside la question de la biodiversité³. Forgé par Arthur Tansley en 1935, il désigne l'unité de base de la vie comme un système d'interaction entre plusieurs éléments, depuis les sols jusqu'aux êtres évolués, qui échangent à travers différents niveaux trophiques de l'énergie ou de la matière⁴. Les mécanismes complexes d'échanges, de transformations et d'interactions au sein d'un biotope naturel tels qu'on les étudie depuis en biologie et en écologie se prêtent logiquement à un traitement sous forme de graphes relationnels (du moins, quand les données sont disponibles). La notion de *biodiversité*, en particulier, relève d'une étude approfondie

2. La revue *Science* a consacré en 2009 un dossier thématique aux réseaux et aux systèmes complexes (325 [5939]).

3. Voir chapitre 12 du présent ouvrage, « Archéologie des connaissances ».

4. <https://fr.wikipedia.org/wiki/%C3%89cosyst%C3%A8me>.

de la notion de système naturel (voire de système socioécologique puisque l'homme en fait lui-même partie) régi par des lois et des équilibres qui le rendent tout à la fois stable et évolutif. Quelques-unes de ces propriétés structurales semblent similaires à celles rencontrées sur le Web : le principe de la pyramide des *foodwebs* proche de celui des couches fonctionnelles du Web, l'organisation distribuée des biotopes qui n'est pas réductible à une arborescence ni à un processus aléatoire (*random*), la structure globale des espèces en interaction où jouent les notions de centralité et de *clusters*, l'orientation et la multidimensionnalité des liens de connexion (relations de *prédation*, mais aussi de *coopération* ou de *symbiose*, de *complémentarité* dans le cas d'espèces pollinisatrices...). L'étude des écosystèmes naturels en biologie et en écologie, tout comme celle d'un écosystème documentaire tel que le Web, fournissent donc quelques arguments de poids pour déterminer sous quel angle, ou sous quelles conditions, peuvent être développés des « écosystèmes d'innovation ». Ce sont exactement ces concepts-là que l'on attend (enfin, ceux qui partagent mon sentiment) quand on invoque l'idée d'écosystèmes d'innovation. Et ce sont ces concepts clés qui faisaient défaut à notre orateur.

Il m'apparaît que résident au cœur même d'une *organisation en écosystème*, comme les écosystèmes naturels, des propriétés absentes (et peut-être incompatibles) de la perspective adoptée par notre orateur dans sa présentation des processus d'innovation lors de la conférence. Indéniablement, les écosystèmes d'innovation (ceux que l'on nous présente comme tels), mériteraient d'être conçus et observés avec les mêmes instruments que ceux déployés en sciences et en biologie de l'environnement ou dans l'exploration du Web, deux domaines qui me semblent partager certaines propriétés communes à titre de système distribué. Cela permettrait d'interroger le concept d'écosystème d'innovation sous différents angles ou à travers différentes questions. Pour le moment, il ne s'agit ici que d'une revue de concepts que je suppose « clés ». Il restera à les éprouver sous la forme d'un dispositif d'observation et d'un travail d'analyse qui relèvent, à mon sens, du domaine émergent de l'*intelligence territoriale*.

La première d'entre elles concerne la *diversité des éléments* ou des acteurs impliqués qui est une caractéristique fondamentale des biotopes, voire du Web lui-même. La question revêt de nombreux aspects, à commencer par le concept générique « d'innovation », un processus qui peut être, tour à tour et sans hiérarchie, *industriel*, *scientifique*, *économique*, *politique* ou *citoyen*. La diversité des modes d'innovation sur un territoire pourrait constituer un indicateur pertinent dans un travail d'observation, voire d'évaluation. De façon conséquente, les modes d'interaction, mais aussi la nature des acteurs entrant en interaction peuvent être posés, depuis le rôle des grands groupes industriels jusqu'aux petits entrepreneurs en passant par le

rôle des ETI (Établissements de taille intermédiaire ou moyennes PME), sans oublier les autres types de stratification avec les établissements publics. Mais ce qui « fait système » réside surtout dans la *diversité des interactions* qui associent plusieurs types d'acteurs. À la diversité des acteurs, on doit donc aussi associer les multiples types de liens qui les connectent, ce que ne sauraient réduire les flèches univoques des schémas linéaires. Sur ce point, il resterait à en dresser la typologie, pour comprendre la nature d'un « écosystème », naturel ou d'innovation, ou pour le susciter comme espérait le démontrer le conférencier. Un écosystème d'innovation supposerait donc des connexions entre divers éléments qui s'effectuent via des liens de nature extrêmement variée (formation, transfert, coopération, compétition, financement, confiance...). La tâche, en termes de description, est immense : la notion de *coopération*, à elle seule, mériterait d'être détaillée minutieusement dans le cadre d'une sociologie des entreprises ou des organisations. C'est aussi l'une des difficultés que l'on rencontre en écologie dans la distribution des liens qui sont d'une grande variété (*prédation, symbiose, fécondation, coévolution, spécialisation réciproque...*). Dans ce qui pourrait devenir un véritable programme ambitieux d'observation des « écosystèmes d'innovation » (ou de ce qui nous est présenté comme tel), il faudrait aussi tenir compte des multiples façons de considérer les liens de connexions, là aussi comme en écologie ou en sciences de l'environnement : je note la relation A->B quand l'espèce A se nourrit de l'espèce B, mais je note B->A quand j'envisage la même relation sous l'angle de l'échange d'énergie où B contribue aux ressources vitales de A. Les solutions d'observation sont multiples, mais visent, fondamentalement, à identifier, et même à évaluer, la multiplicité des relations d'interactions entre des acteurs différents, et, qui plus est, sur un territoire que l'on peut alors assimiler à un biotope.

Toute une grammaire des connexions reste donc à construire pour rendre efficient un dispositif d'observation des écosystèmes d'innovation. Elle permettrait de définir, notamment, les *éléments centraux* d'un écosystème au sens du *web mining* et de la théorie des graphes. La *centralité* n'est pas la *visibilité* et permet de mesurer l'importance d'un élément relativement à sa place ou son rôle pour l'ensemble du réseau⁵. Il en va de même dans les écosystèmes naturels où l'on peut considérer comme centraux les pollens et les abeilles dans la fertilisation végétale, mais comment et selon quels indicateurs identifier de tels éléments dans le champ de l'innovation : s'agit-il des chargés de mission ou des organismes chargés de l'animation d'un réseau comme le sont les pôles de compétitivité ? Les financeurs de l'innovation ou bien les résultats tangibles d'une politique publique ? Rien ne permet de le mesurer actuellement, mais, là aussi, la tâche peut s'avérer difficile, car nous sommes plongés dans un univers où les connexions entre les acteurs comportent

5. Voir chapitre 8 du présent ouvrage, « *Broadcast* ou viralité ? ».

de nombreuses dimensions et, surtout, varient en fonction de leur distribution à l'échelle du réseau (connexions *locales* ou *globales* ?).

L'une des dimensions essentielles d'un écosystème me semble être celle de *l'équilibre des transactions* qui assure à un système sa pérennisation dans le temps et, peut-être donc, sa capacité de résilience (ou sa persévérance à durer malgré les aléas ou les déséquilibres momentanés). Pour le moment, on peut admettre que *l'innovation*, quelle que soit sa nature, naît d'un « écosystème » (à l'image du principe de *biodiversité*), dans lequel elle se nourrit des nombreuses interactions, proches et lointaines, entre une multitude d'acteurs de nature très diverse formant un système qui tend à s'équilibrer à chacune de ses phases d'évolution. En un sens, il faut concevoir l'innovation comme un processus (graduellement ou par crise systémique) de modification de la forme ou la nature elle-même de l'écosystème dans lequel elle prend place. Tout dépend alors de ce que l'on entend par « forme » ou « nature » d'un (éco)système. Le débat est difficile sur ce point, car pour le mener il faudrait disposer d'une capacité à observer les systèmes (naturels, culturels, économiques, des systèmes dits de « connaissances »...) et à valider certaines de leurs dimensions via des données. Mais on doit s'attendre à ce que, par certains aspects, les écosystèmes d'innovation soient en partie régis par une certaine *verticalisation* ou *hiérarchisation* des acteurs ou des activités, par secteurs ou par filières. Tout comme les écosystèmes naturels développent des propriétés de « spécialisation réciproque » des espèces (ou principe local de *coévolution* comme peuvent l'être les filières industrielles régionales) ou d'organisations hiérarchisées avec des *food-webs* dominés par quelques grandes espèces clés (à l'image des grands groupes industriels et des réseaux de sous-traitance). Les regroupements en *clusters* orientés thématiquement (par exemple la logique de déploiement des 8 instituts de recherche technologique sur le territoire ou celle des 71 pôles de compétitivité) semblent en un sens dessiner une organisation pyramidale des projets écosystémiques développés par les acteurs publics avec des spécialisations géographiques et des regroupements de compétences de recherche et de R&D industriels. En ce sens, il faut noter l'alignement presque parfait du côté de la recherche des programmes régionaux, des programmes de l'ANR et du fameux H2020 européen qui revendique une forme de *soft specialization* des territoires européens. Mais, d'un autre côté, il faut aussi s'attendre à ce qu'un écosystème développe des *propriétés structurales dynamiques* qui expliquent leur « agilité » à évoluer, à résister, à faire émerger des *orientations nouvelles et distinctives*. Au risque de grossir les contours, il me semble que ce second modèle d'organisation suppose l'établissement de liens latéraux connectant une diversité d'acteurs et d'activités (les « circuits courts » si l'on veut) et émergents de type *bottom-up*.

Il s'agit là de pointer une série de propriétés que méconnaissent les modèles dominants de l'innovation, car l'on sait que les écosystèmes naturels sont constitués de nombreux liens « transversaux » entre des espèces de natures différentes (richesse de la diversité) et que ces liens expriment une forme d'autorégulation et de diversification des *interactions ascendantes* dans un système et qu'ils visent globalement à le maintenir en équilibre (réciprocité des interactions où chacun gagne et distribue autant). Si l'on suit toujours l'esprit de la comparaison entre écosystème naturel et écosystème d'innovation, cette seconde famille de propriétés doit alerter sur ce qui semble « manquer » aux politiques publiques actuelles : la prise en compte d'une sorte de plasticité structurale où la créativité et l'initiative dépassent toujours les cadres construits *a priori*, les programmes prospectifs. Un écosystème vit aussi de dynamiques de l'*émergence spontanée* de formes d'organisation que l'*on ne sait toujours pas prévoir*, et qui fondent la capacité de réorganisation d'un système, sa résilience ou ses facultés d'adaptation.

On ne peut achever ce premier balayage des propriétés supposées ou attendues d'un écosystème sans mentionner la question des *finalités* poursuivies. Je sais que la question reste large, et difficile à manier, mais elle est indispensable à poser en matière d'innovation et d'économie, et donc de ses incidences sur le modèle *politique* qui les gouvernent. Tout d'abord, il m'apparaît évident que forme(s) d'un réseau d'interaction et finalité(s) du système entretiennent des liens ténus. Sur ce point, on peut supposer que la perpétuation de la vie peut être considérée comme un objectif (fonctionnel) d'un écosystème naturel, comme le montrent ses capacités de résilience, d'évolution, de mutations par-delà les accidents. Mais qu'en est-il d'un écosystème d'innovation ? S'agit-il seulement de « compétitivité », « d'industrie », de « rentabilité » d'un service ou d'une activité nouvelle ? Il m'arrive désormais d'interpeller les jeunes *startups* sur la question du « pourquoi se lancer dans un projet d'entreprise innovante ? » pour réfléchir à la question qui se pose toujours à un moment ou à autre de la finalité des activités de valorisation (ou d'un autre type de finalité), tant il m'apparaît nécessaire d'en définir les *externalités* (réelles ou supposées, positives ou négatives). Et la question est fondamentalement posée à ce qui se présente comme « écosystème d'innovation ». Personnellement, c'est ici que j'aime à comparer écosystèmes naturels et écosystèmes d'innovation et que l'on voit poindre une politique de développement économique et sociétale. En premier lieu, la « valorisation » financière des projets innovants (à travers la création d'entreprises notamment) ne peut constituer la seule finalité d'un écosystème d'innovation, au risque de conduire à des processus très artificiels pour les acteurs comme pour les territoires. Je pointe ici cette *innovation hors-sol* qui consiste à monter une structure porteuse d'un savoir-faire issu d'un laboratoire de recherche (privé ou public), qui n'est qu'une forme de « transfert » ou de valorisation vers des grands groupes industriels, via des fonds d'investissement... le

tout supporté par des aides publiques. Si « l'écosystème de l'innovation est en pleine effervescence [...] et les *start-up* sont un levier de transformation de notre économie » (BPI-France), alors les trajectoires connues des *start-up* semblent incompatibles avec quelques-unes des propriétés attendues d'un écosystème pérenne : soutien de l'État (parfois massif) et aides publiques à des projets qui visent la valorisation à court terme (dont la seule véritable externalité est d'assurer le rendement du capital), *start-up* qui utilisent massivement les stagiaires ou les travailleurs précaires (que l'on fait travailler, notamment, sous le statut dévoyé « d'auto-entrepreneur »), participation à des réseaux, à des associations ou des fédérations qui cloisonnent l'activité innovante de façon verticale sans prendre en compte (au moins dans les critères d'attribution des fonds publics) les multiples externalités qui les lient à leur territoire ou à la société de façon générale. Pour le dire autrement, on doit supposer qu'un véritable écosystème d'innovation s'accompagne d'un modèle explicite de développement économique, dans toute sa complexité... et avec toutes ses externalités.

La pertinence de ces concepts, largement empruntés aux sciences de l'environnement et à la biologie des écosystèmes, dépend naturellement de l'échelle à laquelle on se place. Après tout, on peut parler d'écosystème à différents niveaux d'observation (on peut admettre que dans l'étude du vivant les échelles écosystémiques sont imbriquées et interdépendantes, depuis les sciences de l'environnement et du climat jusqu'à la microbiologie et l'étude de biotopes extrêmement locaux). En matière d'innovation (au sens large) ou de politique d'innovation, les choses se posent à mon sens différemment. En un sens, un écosystème d'innovation s'est développé indéniablement en France sous l'influence des politiques publiques nées du projet de « montée en gamme » des produits et services « compétitifs » dans une « économie mondialisée ». Depuis 2012-2013, nous en avons photographié différents aspects (en particulier avec l'Atelier Iceberg), comme la carte des 8 instituts de recherche technologique en 2014 :

Image 16.1

Réalisée pour l'IRT B<=>com à Rennes lors du premier forum national des IRT en 2013, cette cartographie type « plan de métro » et réalisée par l'Atelier Iceberg rend compte de la façon dont sont associés 28 grands groupes industriels et organismes de recherche aux 8 IRT de France. Dans les deux vues, ce sont les industriels et les organismes de recherche qui font le lien entre les IRT auxquels ils adhèrent. Au niveau national, les IRT et leurs 28 coadhérents semblent former ce qui ressemble à un vaste réseau distribué, multipliant (en théorie) les occasions de coopération. Répartis dans 8 grandes aires métropolitaines (Nantes, Rennes, Lyon, Grenoble, Paris, Lille, Toulouse-Bordeaux et Strasbourg-Metz), ils constituent l'un des rouages essentiels de la *soft spécialisation territoriale* en « *filières d'excellence* ».

Ce type de cartographie renseigne sur le type de spécialisation de chacun des IRT, à la fois thématique et territorial : Railenium dont Lille est le centre de gravité et spécialisé dans les technologies du transport, Jules Verne pour la région nantaise tourné autour des matériaux et de l'énergie, Bcom à Rennes pour les technologies de l'information. Mais la cartographie renseigne aussi sur le rôle majeur joué par quelques grands groupes industriels participant simultanément à plusieurs IRT, un élément probablement très structurant dans l'organisation territoriale nationale née ces dernières années sous l'impulsion des politiques publiques. Ce rôle s'est vu confirmé dans notre travail de cartographie des « technologies clés 2015 », le document programmatique commandé tous les 5 ans par le ministère de l'Industrie et de l'Économie. Il a pour objectif d'identifier les orientations majeures de recherche et développement pour le tissu industriel français. D'un côté, nous avons croisé sous forme de graphe les technologies entre elles pour déterminer les plus centrales et/ou les plus spécialisées. Mais, de surcroît, nous avons identifié plus de 2 000 entreprises nationales ou internationales travaillant sur le territoire français que nous avons ensuite géolocalisées. La frise géographique en bas du poster renseigne sur la distribution géographique des entreprises en fonction de leurs cœurs de métier, une vue générée à partir de la base de données construite pour l'opération.

Image 16.2

Poster « Les technologies clés 2015 ».

Ces quelques indices cartographiques laissent apparaître un type d'organisation qui peut faire penser à un principe *écosystémique* dans lequel les interactions entre les acteurs (essentiellement des grandes structures de recherche publique et des grands industriels) favoriseraient le développement de l'innovation.

Ces exemples illustrent un point capital dans la démarche d'analyse des « écosystèmes d'innovation » : sur quels types d'instruments, de données et de méthodes, peut-on s'appuyer pour identifier des systèmes d'interaction à différentes échelles territoriales ? Comment mesurer des processus comme l'émergence ou « l'innovation ascendante », l'imbrication de divers types de relations (financement, confiance, coopération, complémentarité...), entre des acteurs eux-mêmes divers par nature (agences de développement, laboratoires publics et privés, PME, grandes entreprises, réseaux de sous-traitants, TPE...) ? Et puis, s'il est possible de déployer un tel instrument d'observation et de mesure, à quelles conditions peut-on parler « d'écosystème d'innovation » ? Le principe admet-il des critères ou des degrés, et selon quel modèle politique et économique ? Ces questions restent encore largement d'actualité aujourd'hui, mais soulignent ainsi la nécessité de concevoir un dispositif d'intelligence territorial, intégrant à la fois des données et des modèles d'analyse.

Étant donné la perspective ouverte par la prise en compte de concepts issus de l'écobiologie (diversité des acteurs et des interactions, processus émergents, recherche d'équilibre des transactions, capacité de résilience...), il va sans dire que les indicateurs « macro » de type statistiques, établis pour des comparaisons massives (bases de données produites au niveau national ou européen, de manière systématique par des organismes de traitement d'information qu'ils soient publics – Insee, ministères, OCDE, etc. – ou privés – Bureau van Dijk...) ne suffisent pas. Pour parvenir à observer ou à cartographier des réseaux d'interaction, il faut aussi mobiliser des sources plus « locales » d'information comme les bases où sont archivés les projets financés ou labellisés de coopération entre plusieurs acteurs, les accords de consortium pour exploiter une technologie, les listes des adhérents à différentes structures d'échange ou d'animation collective, identifier les initiatives majeures de codéveloppement qui peuvent réunir des organismes publics et privés, mais aussi des organisations de la société civile et des groupements de citoyens... En somme, concentrer toutes les informations qui permettent d'observer les jeux d'interaction entre des acteurs sur un territoire et ne pas limiter le recueil d'informations à des listes hiérarchisées (par exemple le nombre de brevets) et, qui plus est, réparties selon le découpage administratif des territoires.

Aux informations recueillies, il faut adjoindre un modèle général de traitement des données, et qui peut déboucher sur de premiers essais d'interface ou de

représentation pour accompagner l'analyse. Étant donné la difficulté théorique et méthodologique à définir un écosystème d'innovation, on peut admettre dans un premier temps que nous avons affaire à un objet « multidimensionnel » qui articule (au moins) trois « types » ou trois « couches » : celle du territoire physique ou géographique (et de ses multiples propriétés), celle des acteurs en réseaux ou coopérant (avec des degrés en fonction du nombre et de l'intensité des liens tissés entre eux) et des thèmes, des « contenus » ou des « objectifs » autour desquels ils produisent scientifiquement, techniquement ou socialement. La couche médiane des acteurs coopérants peut se révéler particulièrement intéressante à cartographier pour concentrer les informations et appréhender les processus écosystémiques liés à l'innovation. À ce titre, on peut distinguer au moins quatre grands types d'informations et d'indicateurs (que je produis sous forme de cartes ou de graphes, mais on peut imaginer d'autres solutions) qui permettent d'entamer une description pertinente d'un écosystème d'innovation : la détection des communautés ou des réseaux existant de coopération ou d'affinité (*qui travaille avec qui ?*), le recensement des compétences présentes sur un territoire (*qui travaille sur quoi ? Répertoire de savoir-faire, thématiques de développement, logiques de filières...*), la géolocalisation (*où se situent les acteurs et les thématiques ? Y a-t-il par exemple des phénomènes de concentration locale ?...*) et, enfin, le suivi temporel qui ouvre sur des questions d'évaluation rétrospective ou de scénarisation prospective (*depuis quand ? Selon quel rythme ?*). Si l'on regroupe ces quatre types d'indicateurs dans un espace unique (une interface) et que l'on y associe les informations pertinentes, on obtient alors une première ébauche de ce qui peut devenir un outil d'observation efficace, mais aussi, au-delà, le mobiliser comme instrument d'évaluation et de pilotage de l'innovation en territoire.

Les premiers enseignements de nos expérimentations (notamment avec le projet VisIR en collaboration avec l'Atelier Iceberg⁶) montrent à quel point la question de l'innovation en territoires ne relève pas d'un processus d'agrégation géographique, mais d'une logique de réseau dynamique de coopération et de fertilisation mutuelle entre les acteurs (parfois très différents). Ainsi, ces *entités géothématiques* que sont les pôles de compétitivité peuvent être aussi évalués en fonction du nombre et de l'intensité des liens qui mettent en relation des acteurs, même s'ils sont en nombre réduit, et non pas seulement en fonction de la « visibilité » nationale ou internationale d'un nombre réduit d'acteurs (souvent) les plus importants. Et bien des ensembles géographiques mériteraient d'être ainsi analysés de ce point de vue, en particulier le « plateau de Saclay » que l'on nous présente souvent comme l'avant-garde de l'excellence nationale en matière d'innovation scientifique et industrielle.

6. <https://ateliercartographie.wordpress.com/2011/04/26/visir/>.

Avec une « approche réseau » et des données fiables (produites et filtrées par des acteurs locaux), on peut enrichir les instruments d'observation traditionnels des organisations géothématiques et traiter de questions complexes. Par exemple, qu'en est-il de la réalité des coopérations scientifiques dans le périmètre thématique et géographique d'un IRT, entre des laboratoires d'une université ? Quels sont les foyers de développement technologique et selon quels processus ou quelle logique les acteurs se regroupent-ils ? Par proximité géographique ou par affinités de « métier » ? À quelle distribution (et non pas quelle proportion) géothématique a-t-on affaire entre « petits » et « grands » acteurs ? Où sont les foyers émergents en termes de nouveaux procédés ou de développement de prototypes ? On peut élargir le principe de ce questionnement à des dynamiques qui peuvent être sociales et citoyennes et à toutes les dimensions d'un territoire. Sur ce point, on peut parier sur le fait qu'un tel instrument d'observation des territoires innovants (ou non) ferait apparaître tout le poids de nombreux projets locaux (par exemple autour d'une économie solidaire et de l'écologie). On peut imaginer développer d'ici peu les bases d'un système de veille ou d'intelligence territoriale qui se nourrirait à la fois de données fiables (et non pas de simples statistiques) et de métriques ou d'instruments de mesure tournés autour des concepts hérités, notamment, de l'écologie et de l'étude de la biodiversité. Les niveaux de *zoom* que permettrait un tel dispositif contribueraient, en particulier, à qualifier une dynamique d'innovation de différents points de vue, depuis l'inscription d'un écosystème d'innovation tourné vers la « compétitivité mondiale » ou, à l'opposé, vers l'écodiversité locale.

Les problématiques ouvertes par une démarche d'observation de ce que l'on nous présente comme des *écosystèmes d'innovation* sont nombreuses et diverses. À l'évidence, l'observation n'est qu'un premier pas, le suivant étant l'orientation ou l'aide à la stratégie, voire à l'évaluation. Un point, en particulier, mériterait l'attention dans une telle perspective : le relevé de toutes les formes d'externalité produites (ou que l'on s'attend à voir produites) par un écosystème d'innovation qui ne peuvent se limiter à des questions de valorisation, voire même d'innovation industrielle en négligeant tous les autres aspects sociétaux impliqués (le modèle économique poursuivi en territoires ; les dynamiques sociétales associées, complémentaires, parallèles, voire contradictoires ; les implications en termes d'emplois ; les anciennes et les nouvelles formes organisationnelles générées ou non...). À titre personnel, je suis impatient de voir ainsi analysés les écosystèmes d'innovation dans le domaine du numérique qui, sous forme de *cantines*, d'*incubateurs*, de *pépinières*, de *plateformes d'innovation*, de *quartiers de la création*, de *centres d'innovation* ou de (supposés) *réseaux de start-up* sont censés concentrer les initiatives et multiplier les interactions. Plusieurs dimensions de ces types de dispositifs apparaissent contradictoires avec quelques-uns

des concepts liés aux écosystèmes naturels : concentration géographique du même type d'acteurs, relations faibles avec les territoires environnants, développement de formes précaires de travail, modèles d'innovation hors-sol fortement tournés vers la financiarisation et le marché mondial, trajectoires chaotiques des *start-up* (au mieux, le plus souvent, quand elles sont rachetées), etc. Si l'on tient compte en amont des investissements publics et privés dans ce type de dispositifs et de modèle de développement, il n'est pas sûr qu'à l'analyse on puisse les présenter comme des écosystèmes d'innovation. Le numérique, assurément, représente actuellement un cas d'école en matière d'analyse écosystémique et, selon moi, un domaine à observer... et à évaluer de façon urgente à la lumière de ces premières réflexions.

Postface

Dominique Boullier

Comme le lecteur pourra le constater, la conclusion de cet ouvrage manque. Parce que c'est Franck qui nous manque et qu'il n'a pas terminé cette tâche, somme toute rare chez lui, d'une publication complète. Est-il nécessaire, d'ailleurs, de produire une conclusion à ce récit d'un parcours de recherche, toujours ouvert et toujours en cours, et toujours collectif qui plus est ? On l'aura noté, le « nous » (du collectif, pas de majesté !) est tout aussi présent que le « je » dans ce texte. Cette omniprésence de l'équipe et de la joie du travail d'équipe, souvent sensible dans la narration, et attestée par tous ceux qui ont pu collaborer avec Franck, paraît parfois contradictoire avec cette singularité qu'il revendiquait tout aussi fermement. Sans doute est-ce parce que sa capacité à attirer, à mobiliser dépendait précisément de la singularité forte de sa passion, de sa curiosité, de son désir constant d'exploration qui était tant présent en lui et qui ne l'a pas fait dévier de sa trajectoire. La recherche était tellement chevillée en lui que même le travail de publication, qui tend à clore l'exploration, à produire de la « science faite » selon des normes académiques conventionnelles, lui paraissait parfois inutile, voire insupportable. Le blog s'est révélé ainsi être la forme adaptée de publication, sans souci de validation, de formatage pour un public ou selon des normes, avec le temps et les formats nécessaires à l'explication, au didactisme méthodologique merveilleux que l'on peut constater de nouveau dans cet ouvrage. Au fond, on peut vraiment dire que Franck Ghitalla, dans sa recherche comme dans ses publications, était « un enfant du Web ». Et plus précisément du Web 1.0, celui des liens hypertextes, que du Web 2.0, car les applications et les réseaux sociaux n'ont jamais été son terrain de jeu favori ni même son espace de publication.

Pourtant, ce Web n'était pas inscrit dans sa trajectoire de formation, ni technologique ni scientométrique. Mais sa formation de linguiste structuraliste, que nous avons partagée, est pourtant restée marquante dans sa façon de traiter ces méthodes d'exploration du Web qu'il a développées. Jean Gagnepain avait en effet produit à Rennes un cadre théorique extrêmement formalisé, inspirant et totalisant, sur les compétences humaines, sur les raisons qui nous permettent d'agir de façon à la fois structurée et toujours unique. Dans ce cadre, la raison langagière n'a pas la primauté sur la raison sociale, la raison éthique ni la raison technique qui se combinent toutes, mais doivent être distinguées et disposer de leurs propres concepts. Lorsque Franck Ghitalla vint entreprendre une deuxième thèse avec moi à l'UTC en 1997, il portait en lui cet équilibre des raisons, et notamment cette impérieuse nécessité de prendre au sérieux la technique. Après tout, Compiègne était sans

doute le meilleur endroit où atterrir en France puisque l'objet même de l'université (de technologie) et de l'équipe de recherche en SHS (Costech, Connaissances, organisations et systèmes techniques) était la technique, ses outils, son industrie, sa fabrique, ses usages, et sa dimension constitutive des humains. Sur ce point, en venant tous deux d'une formation comme celle de la théorie de la médiation de Jean Gagnepain, nous étions d'emblée en phase avec ce milieu. Mais alors que l'écriture aurait pu être l'objet préféré d'un linguiste s'intéressant à la technique, Franck avait adopté un angle d'attaque particulier, celui de l'image, autre technique de représentation, possédant ses propres règles. Sa profonde sensibilité à la puissance des images a continué à irriguer ses expérimentations en *dataviz* dont on peut lire un florilège dans cet ouvrage et que ses anciens étudiants ont prolongées, en développant des images de graphes de toutes sortes, au-delà du « plat de spaghetti » qu'il mentionne.

C'est d'ailleurs ce type d'image qu'il m'a présenté en tout premier en 2000 quand, après avoir obtenu son poste de maître de conférences en linguistique, il m'a annoncé travailler désormais sur le Web : une image d'un réseau de boules rouges (car elles étaient en 3D !) censé représenter je ne sais plus quel réseau sur le Web. Développée avec un collègue de Nantes, cette image était supposée emporter la conviction par elle-même et l'on sent encore dans cet ouvrage son émerveillement constant devant la belle image et un certain étonnement devant l'incompréhension ou la demande d'explication du lecteur. Notre commune admiration pour Jacques Bertin et ses principes graphiques facilitait sans doute la traduction pour me convaincre qu'il existait bien « quelque chose », moi qui travaillais sur le numérique et les réseaux depuis près de 20 ans. Mais je restais quand même déstabilisé par cette entrée dans le domaine sous forme d'une carte de « sites », qui prétendait produire du « social », de nœuds qui devenaient aussitôt des « acteurs » et de *clusters* qui devenaient encore plus vite des « communautés » ! Il fallut toute la passion de Franck pour me faire lire Kleinberg, dont il était déjà fan absolu (avant d'y ajouter Watts ou Barabási évidemment), qui me permit de comprendre le lien avec la scientométrie, que je connaissais à travers les travaux de Callon et Courtial et toute la théorie de l'acteur réseau, que Franck ne fréquentait guère. De même, il parvint à me faire accepter (il fallut quelques années !) une approche inductive qui se satisfaisait des approximations que toute la science des données pratiquait depuis longtemps et qu'elle continue de pratiquer avec toute la sophistication des probabilités contemporaines du *machine learning*.

Trois principes donc :

1. la technique est une raison équivalente aux autres ;
2. les images et en particulier les cartes (et donc la *dataviz*) sont des voies d'accès cognitif puissantes à de grands ensembles de données et des vecteurs rapides de conviction ;
3. l'induction, et les *proxies* qu'elle suppose, est une voie de connaissance scientifique créative et valide.

Autant dire trois occasions de percuter les conventions établies dans les SHS ou tout au moins dans une partie d'entre elles. Et de se sentir sans doute définitivement en marge des grands courants, tout en revendiquant ce statut de pionnier, que cet ouvrage devrait contribuer à rendre encore plus évident pour ceux qui ne le connaissaient pas et qui pourtant utilisent des outils ou visualisent des cartes directement inspirées de ses travaux. De ces trois principes, je peux aussi tirer trois remarques qui m'ont toujours frappé dans les travaux et dans mes relations avec Franck.

La technique est une pratique (et non un commentaire sur la technique) et de préférence une pratique artisanale, comme Franck l'indique à plusieurs reprises, en se désignant comme « artisan de haute technologie ». Tout son récit accepte de faire part des bricolages, des « astuces », équivalent ordinaire des approximations dans la science des données sans doute, mais aussi art de « faire avec » les ressources, de calcul par exemple. Et ce fut souvent une préoccupation obsédante pour Franck et ses équipes, obtenir les capacités de calcul nécessaires à leurs projets ambitieux. L'artisan en lui devenait proche de l'artiste quand il pratiquait la mosaïque chez lui, et sans doute avait-il bien perçu la proximité visuelle et technologique entre son art d'assembler des nœuds ou des fragments de pierre ou de céramique (sans parler du fait que le premier navigateur du Web en 1992 s'appelait Mosaïc !!).

Les images de graphes sont un chemin d'accès perceptif privilégié pour voir les structures. En cela, Franck n'a pas dévié de sa formation, fondée sur la détection des structures, des combinatoires, devenues dans ses travaux des *patterns*, terme si souvent utilisé dans cet ouvrage. Cette forme de visualisation du tout, de vue en surplomb, a présenté des avantages politiques évidents par l'effet de « révélation de ce qui était caché » aux yeux des pratiquants d'un domaine scientifique par exemple, mais aussi aux yeux du public quand il s'agit de rendre enfin visibles les stratégies de la secte de l'Église de scientologie ou celles des membres du CAC 40 qui furent parmi les premières visualisations produites au tout début des

années 2000. Je me souviens encore de la délectation avec laquelle Franck nous révélait ces réseaux, avec le sourire de celui qui joue un bon tour aux puissants. Cependant, malgré son positionnement politique personnel affirmé, notamment en faveur de la laïcité, qui lui a sans doute fait choisir certains terrains plus que d'autres, jamais il ne s'est aventuré dans un commentaire politique de ses cartes (alors que le succès de la prédiction du « non » au référendum de 2005 réalisée par ses anciens étudiants de Linkfluence lui aurait permis d'en faire un argument de poids pour des interprétations savantes). Mieux même, tout se passait comme si, confiant dans le pouvoir de révélation de ses cartes, il s'abstenait d'aller plus loin dans l'interprétation. La description modeste, mais robuste des liens entre entités semblait parler d'elle-même. On comprend que cela n'enthousiasmait pas les éditeurs qui, en France plus qu'ailleurs, goûtent avant tout les envolées brillantes qui prétendent diagnostiquer les maux et l'avenir de nos sociétés à partir d'une tête d'épingle empirique, quand Franck préférait accumuler patiemment et artisanalement un tableau précis d'une situation complexe qui, de fait, empêchait de tirer des leçons simples. On peut percevoir cependant dans ce livre que, l'expérience aidant, il se soit senti en mesure de conseiller un peu plus directement le prince que ce soit en matière de stratégies d'innovation (comme le raconte le dernier chapitre) ou en matière de sécurité (avec ses projets – brutalement interrompus – de suivi de l'activité terroriste).

Enfin, l'acceptation de l'induction et des approximations peut sembler, elle, totalement opposée à sa formation de linguiste structuraliste, fondée sur un modèle qui se décline et qui se vérifie, mais qui ne se met que très difficilement à l'épreuve. Pourtant, c'est parce que Franck a toujours été d'une rigueur conceptuelle systématique et d'une clarté de pensée remarquable qu'il a pu arpenter sans dériver un domaine aussi émergent et propice aux fausses promesses et aux discours marketing. Son attachement à la découverte de structures s'est trouvé cependant perturbé par la question du temps, dans le Web comme dans les représentations formelles des graphes. Il en fait part dans tout le chapitre 8 intitulé « *Broadcast* et viralité ? », qui reprend les analyses de réseaux des années 2000 (dont Strogatz notamment). La question de la propagation, des dérivations (dont le *topic drifting* peut être un équivalent sur le plan méthodologique) commençait à le préoccuper. Il trouvait cependant là aussi des « astuces » pour combiner sa vision structurale et son approche inductive, pour concilier l'instantané de la photo du réseau et la nécessité de suivre son évolution, en utilisant par exemple ce terme de « plasticité structurale », tout en reconnaissant l'immensité de la tâche à venir. Nous en avons discuté, souvent à la sauvette, car il savait que c'était mon thème de recherche de prédilection, et je reste avec ce regret profond de ne pas avoir pu mettre en place les collaborations que nous avions imaginées avec tant de joie.

Alors, dans ce domaine pour ce qui me concerne, et dans bien d'autres pour tous ceux qu'il a formés et inspirés, Franck, sache bien que nous continuons l'arpentage, que nous retenons ta leçon de modeste et brillant artisan, que ton œuvre se propage et féconde de nouveaux champs et de nouveaux chercheurs, même si, toi et ton œil affûté et malicieux, vous nous manquez.

Bibliographie des ouvrages de Franck Ghitalla

- F. Ghitalla, *De l'image. Introduction à la théorie de l'image dans la théorie de la médiation*, [Thèse de doctorat en linguistique : sous la direction de Jean Gagnepain], université Rennes II, 1995.
- F. Ghitalla, « De l'histoire à la perspective de l'histoire », *Tétralogiques*, 10, 1996, pp. 21-34.
- F. Ghitalla, « L'hypothèque poétique », dans J. Giot et J.C. Schotte (éds.), *Surdité, différences, écritures. Apports de l'anthropologie clinique*, Bruxelles, De Boeck (Raisonnances), pp. 25-40.
- F. Ghitalla, « Les signes à venir », *Communication et Langages*, 123, Armand Colin, 1998.
- F. Cormerais et F. Ghitalla, « Les nouvelles technologies de la formation et la question des formats », *Revue E.P.I. (Enseignement Public et Informatique)*, 93, mars 1999, pp. 71-81.
- F. Ghitalla, « NTIC et nouvelles formes d'écriture. », *Communication & langages*, 119, 1999, pp. 91-105.
- F. Ghitalla, « L'espace du document numérique », *Communication et langages*, 126, 4^e trimestre 2000, pp. 74-84.
- F. Ghitalla, C. Lenay, « Largeur et profondeur des espaces de compréhension dans l'exploration des réseaux numériques », *VIII^e Congrès de l'ARCO (Association pour la Recherche en Sciences Cognitives)*, Lyon, décembre 2001.
- F. Ghitalla, *Les formats pédagogiques. Théorie expérimentale de l'interaction et de la scénarisation pédagogiques*, U.T.C., novembre 2001.¹
- F. Ghitalla et C. Lenay, « Les territoires de l'information. Navigation et construction des espaces de compréhension sur le Web », *Les Cahiers du numérique*, 3, 2002, pp. 51-63, <https://www.cairn.info/revue-les-cahiers-du-numerique-2002-3-page-51.htm>
- F. Ghitalla (éd.), « La Navigation », *Les Cahiers du Numérique*, Hermès, Meudon, 2002.
- F. Ghitalla, « Arpenter le Web : liens, indices, cartes », *Terminal*, 86, Paris, L'Harmattan, 2002, pp.41-55.
- F. Ghitalla, « L'âge des cartes électroniques : outils graphiques de navigation sur le Web », *Communication et langages*, 131, 1^{er} trimestre 2002, pp. 66-80.²
- S. Bouchardon, F. Ghitalla, « Récit interactif, sens et réflexivité », *Hypertextes et hypermédias. Produits, Outils et Méthodes. H2PTM*, Meudon, Hermès, 2003, pp. 35-46.

1. Rapport de recherche présenté dans le cadre du projet PIEMONT (Plate-forme Intégrée d'Évaluation de Méthodes et d'Outils de formation à base de nouvelles Technologies, 1997-2000) soutenu par le pôle NTE de la région Picardie.

2. Dossier : Nouveaux théâtres de la mémoire.

- D. Boullier, F. Ghitalla, P. Gkouskou, L. Le Douarin, A. Neau, *L'outre-lecture. Manipuler, (s') approprier, interpréter le Web*, Paris, Éditions de la bibliothèque publique d'information/Centre Pompidou, 2003.
- D. Boullier, F. Ghitalla, A. Neau, L. Le Douarin, *Schémas personnels d'interprétation et d'appropriation dans les usages des supports écrits numériques*, Paris, Éditions de la bibliothèque publique d'information/Centre Pompidou, février 2003.
- D. Boullier, F. Ghitalla, P. Gkouskou, L. Le Douarin, A. Neau, « L'expérience du Web : incertitude et diversité des cadres de référence », *Congrès de la SFSIC*, Béziers, juin 2004.
- D. Boullier, F. Ghitalla, « Le Web ou l'utopie d'un espace documentaire », *I3, Information, Interaction, Intelligence*, 4 (1), 2004, pp. 173-189.
- E. Diemert, F. Ghitalla, C. Maussang, F. Pfaender, "TARENTE: an Experimental Tool for Extracting and Exploring Web Aggregates", *Proceedings of IEEE ICTTA04*, Damascus, Syrie, 2004.
- S. Bouchardon, E. Broudoux, O. Deseilligny, F. Ghitalla, *Co-constitution d'un dispositif électronique, d'un champ et d'une communauté*, Paris, Éditions du Centre Georges Pompidou, septembre 2005.
- F. Ghitalla, A. Le Berre, M. Renault, « Des documents, des liens et des acteurs », *Conférence H2PTM 2005*, Meudon, Hermès, 2005.³
- F. Ghitalla, F. Pfaender, C. Maussang, "Extracting and Exploring Aggregates of web Documents with Experimental Visualisation Tools", *Augmented Heritage: New Era for Architectural Design, Civil Engineering and Urban Planning*, Paris, Europa Editions, 2005, pp.57-69.
- F. Ghitalla, « Questions autour de l'archivage du Web », *Les Nouveaux Cahiers de l'Audiovisuel*, 5, 2005, pp.54-57.
- F. Ghitalla, M. Jacomy et F. Pfaender, « Détection et visualisation d'agrégats de documents web. L'exemple du domaine thématique de la Culture Scientifique, Technique et Industrielle », *Rapport du projet WebCSTI*, Cité des Sciences, Janvier 2006
- F. Ghitalla, M. Jacomy, « Analyse topologique du domaine web de la culture scientifique, technique et industrielle », *Rapport final du projet Web CSTI (SP2)*, mars 2007.
- S. Bouchardon, E. Broudoux, O. Deseilligny, F. Ghitalla, *Un laboratoire de littératures – Littérature numérique et Internet.*, Paris, Éditions du Centre Georges Pompidou, 2007.
- D. Diminescu, F. Ghitalla, M. Jacomy, « Méthodologies d'analyse de corpus en sciences humaines à l'aide du Navicrawler », *Rapport final du programme TIC & Migrations*, Paris, Fondation de la Maison des Sciences de l'Homme, 2007.
- F. Ghitalla, « La "Toile Européenne" : Parcours autour d'une cartographie thématique de documents web consacrés au thème de l'Europe et à ses acteurs sur le Web francophone : la carte, un média entre sémiotique et politique. », *Communication & langages*, 158, 2008, pp.61-75.

3. En ligne : <http://www.utc.fr/rtgi/documents/DOC-LIENS-ACTEURS.pdf>

- V. Abajian, F. Ghitalla, S.A. Mahe, O. Matte-Tailliez, "E-Archaeology+: An integrated expert system dedicated to Archaeology", *2008 3rd international conference on information and communication technologies: From theory to applications* , IEEE, 2008, pp. 1-4.

Articles de presse

- « Les opposants au traité européen ont mieux utilisé Internet », article dans *Le Monde* du 9 juillet 2005.
- « Franck Ghitalla, l'as des réseaux », article de David Larousserie dans *Le Monde* du 27 décembre 2012
- « Lumière sur le côté obscur du Web », article de David Larousserie dans *Le Monde* du 5 juin 2017

Articles du blog de Franck Ghitalla (l'Atelier Cartographique)

L'atelier de cartographie. Pratique et enjeux des cartographies thématiques de documents web .	2006
Détection et visualisation d'agrégats de documents web. L'exemple du domaine thématique de la culture scientifique, technique et industrielle.	2006
Réseaux, traces et controverses	2009
La géographie des agrégats de documents sur le Web	2009
Les <i>maps of sciences</i>	2009
Du nuage aux abymes. Dimensions heuristique et expérimentale des modèles du Web .	2009
Converser avec les données numériques. Quelques réflexions, au passage de l'été, sur le « phénomène data » et la « science »	2012
Innover l'innovation	2014
Contribution à la synthèse sur le modèle de l'innovation en France	2014
Les données numériques et l'alchimie quali-quantitative	2014
Cartographier les sciences de l'information et de la communication, les deux premiers clichés.	2014

Table des matières

L'auteur	7
Les éditeurs scientifiques	9
Préface	11
Prologue	15
Chroniques du Web	23
Chapitre 1. Voir le Web	25
Chapitre 2. Questions de distances	35
Chapitre 3. Hiérarchies	49
Chapitre 4. Agrégats	59
Chapitre 5. Méthode et instruments	75
Chapitre 6. Des <i>data</i> ciselées	93
Chapitre 7. Le Web Datarium	109
Chapitre 8. <i>Broadcast</i> ou viralité ?	123
Chapitre 9. L'information, en n dimensions	141
Chapitre 10. La prochaine étape : la multidimensionnalité	157
Chapitre 11. Une première plongée dans le réseau Tor	171
Études cartographiques	187
Chapitre 12. Archéologie des connaissances	189
Chapitre 13. Révélation	207
Chapitre 14. La cartographie d'information ou <i>l'alchimie quali-quantitative</i>	223
Chapitre 15. Cartographier la science	237
Chapitre 16. Les écosystèmes d'innovation	259
Postface	277
Bibliographie des ouvrages de Franck Ghitalla	283

Réalisé sur les presses de l'imprimerie Ciaco
en impression à la demande
depuis le mois de décembre 2020
à Louvain-la-Neuve, Belgique
N° d'imprimeur : 89470