



THATCamp Paris 2012 Non-actes de la non-conférence des humanités numériques

Éditions de la Maison des sciences de l'homme

Évolution de l'archive ouverte HAL-SHS

Proposé par Laurent Capelli

Collectif

DOI : 10.4000/books.editionsmssh.297

Éditeur : Éditions de la Maison des sciences de l'homme

Lieu d'édition : Paris

Année d'édition : 2012

Date de mise en ligne : 1 octobre 2012

Collection : La Non-Collection

ISBN électronique : La Non-Collection



<http://books.openedition.org>

Référence électronique

COLLECTIF. *Évolution de l'archive ouverte HAL-SHS : Proposé par Laurent Capelli* In : *THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques* [en ligne]. Paris : Éditions de la Maison des sciences de l'homme, 2012 (généré le 21 février 2017). Disponible sur Internet : <<http://books.openedition.org/editionsmssh/297>>. ISBN : 9782735115273. DOI : 10.4000/books.editionsmssh.297.

Ce document a été généré automatiquement le 21 février 2017.

Évolution de l'archive ouverte HAL-SHS

Proposé par Laurent Capelli

Collectif

- 1 HAL-SHS est le portail en sciences humaines et sociales de la plate-forme d'archives ouvertes HAL (hyper articles en ligne), créé en 2001. La plate-forme SHS a, pour sa part, été lancée en 2004.
- 2 Le directeur adjoint du Centre pour la Communication Scientifique Directe (CCSD) qui gère et développe HAL depuis le début, Laurent Capelli, présente les évolutions prévues de la plate-forme HAL au printemps 2013, lors de la sortie d'une version 3 complètement réécrites. Comme toute application, HAL a vieilli, même si le code PHP a été retouché au fur et à mesure. Il était donc temps de faire une nouvelle écriture de l'application. Cette réécriture ne sera pas fondée sur un dépôt d'archives existant (type DSpace), mais sur le même cœur (*core*) que Sciencesconf.org, développé ces dernières années par le CCSD en parallèle avec HAL. Les évolutions les plus importantes concernent l'élargissement des types de documents déposables aux vidéos, la création d'un moteur de recherche, la possibilité de créer une page chercheur. Ces évolutions mènent à des interrogations plus globales sur le statut des documents déposés dans une archive ouverte et sur l'usage (détourné ?) de la plate-forme à des fins d'évaluation.

Intégration d'un moteur de recherche

- 3 Actuellement, il n'y a pas de moteur de recherche associé à HAL, au sens d'une indexation des métadonnées et des données déposées. En fait, le moteur actuel de HAL effectue de simples requêtes SQL dans la base de données MySQL, sans entrer dans le contenu des documents déposés. Il est donc prévu de développer un véritable moteur de recherche à facettes. Le software n'est pas encore choisi, entre AFS (une des briques d'Isidore) et Solr (moteur de recherche Open Source). Le futur moteur de recherche interrogera également

les données Exif des images et des vidéos. Il y aura des limites dans la recherche en plein texte dans les documents textuels, dans la mesure où les pdf ne sont pas structurés.

Élargissement de la typologie de documents déposables aux vidéos

- 4 La nouvelle version de HAL sortira au printemps 2013, sous la forme d'un nouveau portail dans lequel on pourra directement déposer tous les types de documents. La grande nouveauté est l'extension du type de documents déposables aux données de la recherche et en particulier à la vidéo, après l'expérience réussie de l'élargissement aux images (à partir de 2008, portail MediHAL créé en 2010). HALvidéo est développé dans le cadre de la v3 et sera intégré dans la nouvelle plateforme. La vidéo pose un problème spécifique de stockage, les fichiers vidéos étant beaucoup plus volumineux que des fichiers texte, PDF ou même image. Actuellement, le CCSD a la capacité d'héberger 5 000 heures de vidéos sur ses serveurs, mais cette capacité de stockage n'est pas limitée et peut être augmentée - rappelons que le CCSD se situe dans les locaux de l'IN2P3. 5 000 heures de vidéos correspondent à une journée d'enregistrement du LHC. Même si au départ le CCSD ne disposait pas de penta octets pour accueillir des dizaines de milliers d'heures de vidéos, le ministère a assuré mettre les moyens pour un portail vidéo conséquent.
- 5 Différents types de vidéos pourront être déposés : enregistrement d'une conférence ou d'un colloque, vidéo expliquant un article, etc.
- 6 Beaucoup de centres de recherche ont accumulé des heures de vidéos et les ont hébergées de manière disparate. Par exemple, le Centre Edgar Morin, qui a une cellule audiovisuelle depuis 20 ans, a déposé une partie de ses vidéos sur canal-u, une partie sur le réseau public parisien. Ce corpus a-t-il vocation à être hébergé sur HALvidéo ? Quels types de vidéos peuvent être considérés comme des archives ? On peut en effet distinguer divers types de documents et questionner leur caractère scientifique.
- 7 Pour saisir les enjeux liés au dépôt de vidéos, on peut se rapporter aux pratiques déjà en place avec la photographie, les images fixes. Par exemple, quel est le statut d'une reproduction numérique d'un manuscrit médiéval réalisée à l'IRHT ? S'agit-il d'archive de la recherche ou d'archive tout court ? Autre exemple, l'archéologue revient du terrain de fouille avec des photos : il peut les confier à son laboratoire ou les déposer dans MédiHAL.
- 8 La plateforme HAL n'est pas un dépôt légal, il y a une autre institution pour cela. Le dépôt sur HAL peut tout à fait coexister avec le dépôt légal.
- 9 On peut prendre l'exemple de MédiHAL, qui est utilisé par des laboratoires comme un complément à une plate-forme existante comme une photothèque internet mais ne gérant pas l'archivage ou la diffusion. MédiHAL sert alors à la fois à assurer l'archivage à long terme et à mieux diffuser les données. Dans d'autres cas, c'est l'ensemble de la photothèque qui a été basculée dans MédiHAL parce que le laboratoire n'arrivait plus à maintenir son application.
- 10 HALvidéo permettra ces deux usages : le dépôt individuel de vidéos liées à la recherche, comme source ou comme explication mise en scène, de conférences et colloques, mais aussi le dépôt de collections de vidéos par des laboratoires ou des institutions de recherche

- 11 Dès l'origine, HAL a été pensé pour accueillir des articles et les données liées à cet article, en permettant le dépôt d'annexes. Dans la nouvelle plate-forme, il sera possible de faire des liens entre des productions scientifiques et des documents, des images, des séquences audiovisuelles.
- 12 HAL, pour aucun type de document, n'est un outil de traitement ou d'analyse des données. Il est seulement un outil de dépôt et de diffusion des données. Il stocke des documents avec des métadonnées minimales. Toute analyse ne peut se faire que dans une application tierce qui se branche sur HAL via une API. Cette API, qui existe déjà, sera également améliorée dans la version 3 de la plate-forme. Les métadonnées n'ont pas à être polluées par telle analyse d'untel à tel moment. De même, en ce qui concerne la vidéo, HAL fournira un outil de diffusion des vidéos : une fonctionnalité simple de chapitrage et d'annotation minimale des vidéos et la possibilité d'incorporer (*embedder*) la vidéo sur une page web (fonction qui sera automatisée pour les carnets de la plate-forme Hypotheses.org, en insérant juste l'identifiant de la vidéo, un lecteur sera affiché).
- 13 HAL fonctionne ainsi comme un gros disque dur virtualisé quelque part, que l'on peut interroger via une API, un protocole comme l'OAI-PMH, et diffuser via un flux RSS ouvert et personnalisable.
- 14 L'élargissement de la typologie des documents aux vidéos pose aussi la question d'autres types de documents : rapports ANR, mémoires de recherche, rapports de fouilles, articles de vulgarisation, billets de blog.

Le référentiel chercheur et la question de l'évaluation

- 15 Une autre fonctionnalité, demandée depuis longtemps par certains utilisateurs de la plate-forme, sera implémentée : l'ajout d'un référentiel chercheur permettra de rassembler les publications de l'auteur sur une page sur laquelle le chercheur pourra également ajouter son CV, sa bibliographie et où il pourra afficher quelques éléments « statistiques » sous la forme d'un graphe : avec qui il publie, de quels pays sont les co-auteurs, dans quelles revues, dans quelles catégories ou disciplines, à quel rythme. Ces éléments ne seront affichés sur la page du chercheur que s'il le souhaite. Ces statistiques ne valent évidemment que pour les documents effectivement déposés dans l'archive HAL.
- 16 Mais la question se pose de l'usage possible qui peut être fait de ces statistiques du point de vue de l'évaluation de la recherche et des chercheurs. On peut pointer l'incomplétude des données en question, critiquer le fait que HAL puisse être utilisé pour l'évaluation de la production scientifique des chercheurs alors qu'il n'est pas fait pour cela. Le simple fait de donner la possibilité d'afficher ces données sous la forme d'un graphe sur la page personnelle du chercheur met en lumière un usage que peuvent faire les instances d'évaluation d'une archive ouverte. On peut cependant remarquer que lesdites données existent de fait dans un dépôt d'archives ouvertes et qu'il suffit de les extraire et de les exploiter, ce que certaines instances font déjà. Même si HAL est un outil d'auto-archivage et non d'évaluation, il est de fait utilisé dans ce sens. Dans un climat général où les instances sont à la recherche de tous les indicateurs et de tous les chiffres possibles, HAL est bien une source de données, parmi d'autres, servant à l'évaluation.
- 17 D'ailleurs, dès la création de HAL, les laboratoires ont demandé à pouvoir afficher les productions de leurs chercheurs. Dès 2000, la demande du CNRS portait sur l'estimation du nombre de publications du CNRS dans arXiv.org, l'archive des physiciens. Il fallait

donc pouvoir afficher les publications de chaque institution, générer des listes bibliographiques. Par la suite, certains établissements ont mis en œuvre une politique volontariste d'incitation voire d'obligation destinée aux chercheurs afin qu'ils déposent l'intégralité de leur production dans HAL. C'est le cas, par exemple, de l'université de Liège ou de l'Institut national de recherche en informatique et en automatique (INRIA, France), dont les chercheurs seront poussés à déposer tous leurs articles à partir de 2013 sur son site propre sur HAL. Concrètement, un chercheur de l'INRIA ne sera pas autorisé à publier un article si son *preprint* n'est pas déposé au préalable dans HAL. Pour conclure sur ce point, il semble important d'indiquer que HAL n'est pas un outil d'évaluation, mais un réservoir d'articles scientifiques décrits de la façon la plus correcte. Concernant les statistiques affichées pour les dépôts, on peut suggérer certains affinements et un ajout : lors de la consultation des fiches, il faudrait que le passage des robots soit différencié des visites ; et qu'une statistique du moissonnage de l'article via le protocole OAI-PMH soit affichée, permettant de savoir dans combien de bases, voire dans lesquelles, l'article a été moissonné et est donc présent.

Dépôt du texte intégral et de notices simples

- 18 Le dépôt d'un article sur une archive ouverte comme HAL n'est pas sans poser de difficultés dans certaines disciplines, en particulier vis-à-vis de certains éditeurs et revues de SHS qui n'acceptent pas d'articles ayant déjà été déposés dans HAL. De ce point de vue, le rapport de force actuel, en particulier avec les éditeurs privés de revues, n'est pas forcément à l'avantage des chercheurs. On peut cependant inciter les chercheurs à négocier avec leurs éditeurs la possibilité de déposer la version d'auteur dans une archive ouverte comme HAL. La question demeure de la manière de valoriser pour le chercheur également les notices correspondant à des articles qui n'ont pas pu être déposées par l'éditeur.

Épisciences.org : création d'épi-revues

- 19 Il est également prévu, dans la nouvelle version de HAL, de pouvoir créer des épi-revues qui se distinguent des simples collections existantes. Un groupe scientifique pourra constituer un comité de lecture qui sélectionnera des papiers déposés dans le réservoir, les évaluera et proposera à leur(s) auteur(s) de les publier dans ladite épi-revue.
- 20 Une épi-revue se distingue d'une revue à qui les articles sont soumis, via un appel à contribution ou de manière volontaire. Dans le cas d'une épi-revue c'est en effet le comité de rédaction qui effectue une veille ou une recherche sur ce qui est effectivement déposé dans l'archive ouverte et qui va proposer à l'auteur d'intégrer son article dans l'épi-revue en question. Dans d'autres cas, un appel à contribution classique peut être lancé et la soumission des articles proposés s'effectue alors simplement par un dépôt en archive ouverte. Un travail éditorial (sur le fond et la forme) va alors être réalisé sur l'article au format papier, et l'auteur pourra déposer une version 2 – la version 1 étant conservée dans HAL suivant le principe qu'aucun document déposé ne peut être retiré. On rappelle que le CCSD ne valide pas les articles déposés après évaluation scientifique, mais seulement après évaluation sur le plan technique (format, fichier, métadonnées, type de document). L'idée d'une épi-revue est donc de faire quelque chose, à un niveau

scientifique, des contenus déposés dans le réservoir. Cela ne revient donc pas au même que d'effectuer un tamponnage des articles, puisqu'il y a un travail de sélection scientifique.

- 21 Cette fonctionnalité est née d'une demande, d'un besoin de créer des journaux scientifiques en validant des articles déposés dans HAL – ou dans un autre réservoir. Il faut noter qu'ArXiv.org, qui a inspiré la création de HAL, n'a pas évolué vers ces nouvelles fonctionnalités.
- 22 Ce nouvel objet éditorial a répondu à une demande de chercheurs de certaines disciplines, en particulier en mathématiques et en physique alors que HAL n'en était qu'à ses débuts. L'INRIA y a joint sa propre demande. Les sciences humaines et sociales sont moins concernées par cette fonctionnalité, hormis la revue TAL (Traitement automatique des langues), qui utilise d'ores et déjà de manière détournée Sciencesconf.org pour gérer l'évaluation des articles. Cette demande vient de revues hors SHS qui ne disposent pas d'une plateforme de diffusion électronique telle que Revues.org destinée aux revues en SHS.
- 23 Il s'agit donc d'un nouvel outil développé par le CCSD, Episciences.org, qui sera connecté à HAL et permettra de développer des épi-revues. Son *workflow* se calera sur les pratiques courantes d'évaluation des publications scientifiques. Il sera donc proche de celui sur lequel repose OJS (Open Journal Systems) utilisé par le service de Manuscrits proposé par Revues.org. Il permettra de nommer des lecteurs externes, de créer une grille de lecture, d'échanger avec l'auteur qui pourra faire une version 2 ou 3 modifiée et la déposer sur HAL, de créer un numéro, etc.
- 24 Pour aller plus loin, il est possible de consulter un article sur la question des épi-revues à partir des archives ouvertes comme HAL, « De la numérisation des revues à leur déconstruction numérique »¹. Cette initiative peut être rattachée au Budapest Open Access Initiative (BOAI) de 2001 : mercredi 13 février 2002, un collectif de chercheurs a déclaré la guerre aux revues scientifiques éditées par des « grands éditeurs » comme Springer-Verlag ou Elsevier. La création de revues alternatives, en ligne et gratuites, est pensée comme un moyen pour empêcher les grands éditeurs de prendre à terme le contrôle de la production de la littérature grise. Ce débat reste d'actualité en 2012, avec d'une part, le lancement d'une pétition de chercheurs appelant à ne plus soumettre d'articles aux revues de Springer ou d'Elsevier et d'autre part, la multiplication d'appels à la publication des résultats de la recherche en open access.

Conclusion

- 25 Pour conclure, cette nouvelle version de HAL comprend un élargissement de la typologie des documents déposables, avec l'introduction de la vidéo (et des services afférents : chapitrage et lecteur exportable), d'un véritable moteur de recherche, et un développement des fonctionnalités (référentiel chercheur et pages personnelles, Episciences.org). Ces élargissements suscitent des interrogations légitimes sur la frontière entre les documents recevables et ceux qui ne le sont pas - ce qui est particulièrement aigu pour la vidéo où les fichiers sont lourds et pour les questions de droits à l'image - et sur les usages de la plate-forme pour l'évaluation, une demande émanant des institutions, mais soumise à la critique par les chercheurs.

NOTES

1. Jean-Max Noyer, Gabriel Gallezot, Olivier Ertzscheid, Ghislaine Chartron, “De la numérisation des revues à leur déconstruction numérique”, *Les Cahiers de la SFSIC*, n°2, avril 2008, p.15-17.

RÉSUMÉS

Le CCSD est en train de repenser son archive ouverte HAL. Il propose d'en dévoiler les futurs contours : le chercheur au centre du dispositif, moteur de recherche, dépôt de vidéo, épirevues, etc.

INDEX

Mots-clés : archives ouvertes, HAL, HAL-SHS, Open Archives, vidéo, identité du chercheur, évaluation