Connecting the Knowledge Commons — From Projects to Sustainable Infrastructure

Leslie Chan and Pierre Mounier (dir.)

DOI: 10.4000/books.oep.8999
Publisher: OpenEdition Press
Place of publication: Marseille
Year of publication: 2019
Published on OpenEdition Books: 2 June 2019
Serie: Laboratoire d'idées
Electronic ISBN: 9791036538025

Electronic reference

© OpenEdition Press, 2019
Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International - CC BY-NC-ND 4.0
CONNECTING THE KNOWLEDGE COMMONS
FROM PROJECTS TO SUSTAINABLE INFRASTRUCTURE
CONNECTING THE KNOWLEDGE COMMONS
FROM PROJECTS TO SUSTAINABLE INFRASTRUCTURE

The 22nd International Conference on Electronic Publishing - Revised Selected Papers

Edited by Leslie Chan and Pierre Mounier
Acknowledgements

The editors would like to thank their colleagues who took an active part in reviewing and editing the papers included in this book: Serge Bauin, Jen Booth, Jean-Claude Guédon, Delfim Leão, Panayiota Polydoratou, Laurent Romary, Joachim Schöpfel, Jadranka Stojanovski.

They would also like to thank the OpenEdition Press team: Gwenn Dossmann, Cédric Gaultier and Amandine Texier.
Introduction

Open Infrastructure: From Monocultures to Bibliodiversity

Leslie Chan
University of Toronto Scarborough
chan@utsc.utoronto.ca

This volume brings together selected full papers from the 2018 ELPUB conference on the theme of *Connecting the Knowledge Commons: From Projects to Sustainable Infrastructure*. These papers went through a second round of external review after the conference, then they were revised by the authors accordingly.

ELPUB, which stands for Electronic Publishing, is one of the longest running international conference series dedicated to exploring and sharing ideas on the impact of new information and communication technologies on the practice of “publishing” and information dissemination. Over the years, ELPUB has brought together a diverse international community of librarians, developers, publishers, entrepreneurs, administrators, policy makers, and researchers across the disciplines in the sciences and the humanities.

With the first conference held in 1997, the earlier iterations of the series had a rather broad base attention to publishing across various cultural and industrial sectors. In recent years however, the focus of the series has been more concentrated on academic or scholarly communication in particular. Not surprisingly, Open Access (OA) and its associated technology and social support systems have become one of the key topics of the conference series in recent years.

The conference series therefore reflect and document the many major changes taking place in the scholarly communication landscape over the last two decades. While it appears that the moral battle for Open Access for the public good had largely been won, the battle over how best to provide and sustain OA continues to be hotly

---

1. For a brief history on the origin of ELPUB, read Olivia Holmes essay here: https://elpub2019.hypotheses.org/a-tribute-to-john-smith
2. All the Proceedings of ELPUB since its inception are archived at the Computer Science Bibliography database https://dblp.org/db/conf/elpub/index.html and at the ELPUB Digital Library https://elpub.architexturez.net/
contested. But one thing remains constant. The object of OA continues to be the journal and its content, while maintaining prestige and garnishing metrics continue to be the key incentive for researchers and their institutions. This has allowed legacy publishers with their familiar brands to continue with their strong hold on the reputation economy.

Innovations, such as the so-called megajournals pioneered by PLoS and founded by researchers, have led to interesting developments in peer review, article level metrics, and data sharing. It had also helped expand the conservation on OA to Open Science, the umbrella term that calls for opening access to the entire research life cycle. This in theory had the potential to further democratize access and participation in the knowledge production process. But the exclusive reliance on the APC model and its success in generating handsome revenue has created such a boom in the industry that many new to OA have often mistaken the APC model as synonymous to OA. The replication of the megajournal model by new and legacy publishers has driven the debate further towards the economic aspects of OA, focusing on who pays, how much, and how “sustainable” such payments would be. In other words, sustainability has become how much money the current system, largely supported by public funding, could continue to pay the private sector publishers.

The announcement of PlanS last year only amplified the economic dimension of the issue. It is primarily designed to control the rising costs of APC, but it will unlikely disrupt the existing and indeed growing power structure of the publishing industry. At the same time, it will likely lead to an over-representation of authors who have the advantage of receiving funding from funders who mandate OA, while unfunded researchers and those in poorly resourced institutions will be under-represented under the APC model.3 This has particularly strong and negative implications for researchers and institutions from the global South. And colleagues from Latin America, representing community based projects and the new initiative AmeliCA (Open Knowledge for Latin America and the Global South) have been articulating forcefully why the global North model will further disadvantage scholarship from the global South.4

The inequity being amplified by the APC model forces us to rethink the meaning of openness, and why we favor OA in the first place. OA is a means to various ends, and these ends are dependent on the context of the knowledge producer and the user. Policies like PlanS are focusing on the visible problems of accessibility by trying

---

3. Studies documenting such outcomes are beginning to emerge. See for example Beasley and the recent study by Khoo.
4. See http://www.amelica.org/ for details on the potential negative consequences of policies and standards from the global North on other parts of the world, and why locally driven community based initiatives are far more appropriate for local needs.
to eliminate the paywall, but it fails to address the embedded structural problems that are deeply rooted in the knowledge production system. Understanding these often hidden and invisible structural problems requires us to examine the history and political ideology that form the foundation of the western knowledge system, which favors hierarchies of knowledge, while privileging the normative science and scholarship from the West as the standard by which all other forms of knowledge are being measured. The dreaded, or revered, journal impact factor is but a visible form of “standard” and power that enforces a set of invisible norms that only those privileged enough to gain access to them are given additional advantage to succeed. On the other hand, knowledge outside of the norm are often rendered invisible by this system of power.

The consolidation of the legacy power teaches us an important lesson. Despite the sustained interest, the OA movement and OA communities are really a hodge-podge and at times loosely connected organizations, coalitions, and individuals that have very different motivations and incentives for achieving OA. As a result, they often work at cross-purposes. In the meantime, the commercial interests are unified in their goal, that is how to maximize their growth, create new revenue streams, and sustain their profit margins. They have clear sustainability plans, the OA communities do not, and they have successfully co-opted the OA agenda in their favor.

In recent years, some of the well-resourced multinational publishers like Elsevier have been busy transforming themselves from content producers to “analytics” companies, building end-to-end platforms through merger and acquisitions. They understand that in the growing platform economy, big data on institutional performance and productivity in research are the “new oil” waiting to be extracted and turned into new financial products. As higher education becomes highly internationalized and marketized, competition to maintain or improve on their global rankings are becoming imperative. As a result, universities are willing not only to pay substantial amounts to buy the analytics and “intelligence” that would inform their decision making, they are also willingly giving away their institutional data while outsourcing their infrastructure needs to the same publishers to ensure their competitiveness in the global arena.

The OA communities are awakening to this growing trend of the vertical integration of infrastructure, and organizations such as SPARC and others have been raising the alarm bell about the potential consequences when publishers not only have near monopoly power over pricing, but also begin to take “soft” control over university’s decision making processes on resource allocation on research based on big data analytics. This is a far bigger concern than the one raised by Bilder, Lin

5. See the SPARC landscape analysis: https://sparcopen.org/our-work/landscape-analysis/
and Neylon that “Everything we have gained by opening content and data will be under threat if we allow the enclosure of scholarly infrastructures.”

It was against such a backdrop that the 2018 and 2019 ELPUB conferences were conceived. The ELPUB venue allows a broad community to think collectively about what constitutes scholarly infrastructure, and how to ensure that infrastructure are kept “open” and under community control and governance. Of course one of the persistent challenges has always been funding and the sustainability of social systems that underlie the physical and technical infrastructure.

There are also growing lessons that simply focusing on the technical and economic dimensions of OA and infrastructure are not sufficient and we have to pay equal if not more attention to the socio-political dimensions of infrastructure design and building. We need to think beyond whether the system is open or closed, but also who builds them, who sets the agenda, who makes the decisions about standards, for what purpose, and who has control and governance of the system. As infrastructures are never neutral, we need to be cognizant about biases that may further entrench inequity in whose knowledge is privileged and whose knowledge is made invisible by the current system.

In this regard, the notion of “bibliodiversity,” the theme of ELPUB2019, is particularly pertinent. A term coined by Chilean publishers in the 1990s:

Bibliodiversity is cultural diversity applied to the world of books. Echoing biodiversity, it refers to the critical diversity of products (books, scripts, eBooks, apps and oral literature) made available to readers. Bibliodiversity is a complex self-sustaining system of storytelling, writing, publishing and other kinds of production of oral and written literature. The writers and producers are comparable to the inhabitants of an ecosystem. Bibliodiversity contributes to a thriving life of culture and a healthy eco-social system. (L’Alliance internationale des éditeurs indépendants)

Far from promoting bibliodiversity, the dominant “ecosystem” today has increasingly been prescribing what Vandana Shiva has called the “monocultures of the mind”, as the logic of the system has been driven primarily by market imperatives. In this regard, sustainability cannot simply be about financial viability. Instead, it should be about enabling a diversity of systems that “preserve and strengthen plurality and the diffusion of ideas” and allow for the participation of diverse actors and institutions.

So how should we co-create a healthy ecosystem of knowledges that celebrates diversity of thoughts while addressing epistemic injustice? This remains the biggest challenge for the Open Access communities.

References


---

7. The idea that knowledge practices and institutions may be structured and enacted in ways that simultaneously privilege certain epistemic values (such as universality, objectivity, and truth) while being unjust or dismissive towards particular knowers or ways of knowing (Fricker; Santos).
Abstract: This paper attempts to illustrate the implications of a simultaneous redirection of the big publishers’ business strategy towards open access business models and the acquisition of scholarly infrastructure utilizing the conceptual framework of rent-seeking theory. To document such a transformation, we utilized financial databases to analyze the mergers and acquisitions of the top publicly traded academic publishers. We then performed a service analysis to situate the acquisitions of publishers within the knowledge and education life-cycles, illustrating what we term to be their vertical integration within their respective expansion target life-cycles. The vertical integration is analyzed via a rent theory framework and described to be a form of rent-seeking complementary to the redirection of business strategies to open access. Finally, the vertical integration is noted to generate exclusionary effects upon researchers/institutions in the global south.

Keywords: knowledge production, rent-seeking, scholarly infrastructure, concentration, inequality.

Introduction

As major commercial academic publishers’ have redirected their business strategies to open access and alternative paying structures, it could be argued that this represents a move towards more democratic access to knowledge. However, this paper
problematizes this claim by documenting and examining the simultaneous redirection of multinational academic publishers’ business strategy towards the acquisition and integration of scholarly infrastructure, the tools and services that underpin the scholarly research life cycle, many of which are also geared towards data analytics for the purpose of creating new income streams. We argue that moves toward openness and increased control of scholarly infrastructure are simultaneous processes of rent-seeking which could further entrench publishers’ power and exacerbate the vulnerability of already marginalized researchers and institutions.

This paper’s primary objective is to systematically document the economic concentration in the academic publishing industry in order to situate recent shifts towards the acquisition of scholarly infrastructures and evaluate its implications for inequality in knowledge production. We hypothesize that recent shifts have been facilitated by an already disproportionate publisher ownership of academic content (and data) and that the implications include increased dependence by individual researchers and institutions, as well as the consolidation of an already unequal scholarly communications landscape, making it harder for alternative services and products to succeed in the industry. This paper examines this claim by undertaking an empirical case study of the mergers and acquisitions of three of the five major academic publishing companies (RELX-Elsevier, John Wiley & Sons-Wiley Blackwell, Informa-Taylor Francis).

In particular, the study focuses primarily on the case of Elsevier and Wiley as the companies present a longer history of transitioning towards infrastructure and data analytics. We recognize that the inclusion of Digital Science - Springer-Nature would be particularly timely given recent moves by the company. However, the methodology we employ requires publicly listed data which Springer-Nature does not provide as a private company. We encourage future research which employs a methodology that permits the analysis of merger and acquisitions by Springer-Nature.

Conceptual Framework

This paper builds upon previous work which has empirically documented and questioned the oligopolistic tendencies within the academic publishing industry following the digital turn (Larivière et al.). The work of Larivière et al. demonstrates an increasingly disproportionate ownership of academic journals and papers in the natural and the social sciences by the top 5 academic publishers (Elsevier, Wiley-Blackwell, Springer, and Taylor Francis and Sage). Their findings suggest that the top five publishers accounted for more than 50% of all papers published in 2013 (Larivière et al.). We build upon their findings by documenting a simultaneous
process of concentration of scholarly infrastructure within academic publishing and higher education and analyzing its contribution to the oligopolistic tendencies in the industry.

We propose that the process of concentration of content and infrastructure by academic publishers are inherently interconnected so that they should be analyzed together when looking at consolidations of big publishers’ control and their movement towards open access. The analysis follows a political economy theoretical framework, and draws on the conceptual frameworks of rent-seeking and value grabbing in its analysis of the evolving business strategy of big academic publishers.

These analytical tools facilitate a deeper understanding of how rent relations are asserted and maintained and their implications for inequality in scholarly communications. The relevance of rentiership theory to the political economy study of the Science and Technology Studies (STS) field has been outlined by Birch. Originally conceived as relating to the ownership of and value extraction from land (Ricardo; Marx), the analysis of rent relations has expanded to include natural resources such as oil and also patents, culture and STS (Bina; Harvey, “The Art of Rent”; Birch). Such expansions are particularly relevant to studying the information economy and the role of academic publishers within it (Schwartz).

We focus on two particular transitions within the publishing industry to elucidate the way in which rent-seeking theory can shed light on the rationale guiding publishers behaviour as well as on the potential implications of such behaviours. The first is the transition to the digital era which completely restructured the costs of reproduction of academic content, theoretically decreasing costs and allowing for knowledge to be more accessible. We focus upon the implications of such transition to the profit structure of academic publishers and the subsequent production of the paywall which maintained publishers’ profits. The second transition is one towards the concentration of scholarly infrastructure and services which we document throughout this paper.

Economic rent is the value which can be extracted through the ownership and control of a particular resource (Ricardo). As such, rent relations are social relations between the owners of an asset and the general population which wants access to that asset (Ball). This rent relations have led to the rise of rentier class for which property of an asset represents significant financial security with little involvement on the production of the asset (Veblen; Tawney; Keynes). As such, rent relations are social relations of value distribution that the owner of an asset enjoys as a result of the ownership of the asset rather than of its production. Value grabbing is defined as the appropriation
of surplus value through rent relations (Andreucci et al.). In this respect, value grabbing is an important analytical tool through which to visualize the politics involved in the distributional relations generated by economic rent (Andreucci et al.). Rent struggles are in turn representative of struggles over the redistribution of value that has been grabbed through rent-seeking strategies (Andreucci et al.). Throughout the analysis we build upon the rent-seeking strategies through which economic rent is acquired and maintained in order to demonstrate the extent to which they align with the behaviours of big academic publishers.

In documenting publishers rent-seeking strategies we rely upon concepts such as accumulation by dispossession, monopoly rents and club-goods (Buchanan; Schwartz; Andreucci et al.; Birch). An essential condition for the acquisition of rent over an asset is the entitlement or property title over that asset as established by regulatory and institutionalized property regimes (Birch). Accumulation by dispossession refers to the establishment of such a private property rights for assets that did not previously form part of the private - public ownership regimes (Harvey, “The “New” Imperialism”; Andreucci et al.). In this regard, establishing property titles over an asset such as knowledge requires a process of turning it into a “thing” that can be packaged, turning it into an asset that can be valued, and finally extracting value or rent over that asset (Birch).

The consolidation of academic knowledge in the form of an academic paper which can be owned and distributed for the extraction of profits is an example of the aforementioned process. However, the extraction of profits of academic papers was complicated by the turn of the digital era which enabled a near zero marginal costs for their reproduction. In this context, the establishment of the paywall was a rent-seeking strategy geared at establishing monopoly rent in response to the threats of the digital era. The extraction of monopoly rent is the “result of a systematic shortage of supply created by the property monopoly of the supplier of a key product, which encounters no direct competition from substitution goods” (Zeller 98). The paywall generated a systematic and artificial shortage of supply of academic papers, which in many instances have no substitution, which has allowed publishers to maintain monopoly rent over academic content throughout the digital era.

Rent-seeking theory explains why the turn of the digital era allowed the publishers to establish themselves as a rentier class that freely acquires academic papers as intellectual property assets and then captures rent through their entitlement. Club goods theory provides a framework through which to understand how knowledge in the form of academic papers in the digital era was converted into goods which are non-rival, such that the consumption of the good by one does not impede the
consumption by another, yet excludable, so that one may be excluded from its consumption through artificial scarcity (Schwartz; Hartley et al.).

Given that rent is a distributional and not a productive relationship it has the potential of draining productivity by limiting the productive potential of members who don’t have access to the assets as a result of the artificial scarcity (Harvey, *The Limits to Capital*; Felli; Andreucci et al.). As a result, the expansion and proliferation of rent based returns has also been associated with exacerbating inequality over the past decades (Piketty). Furthermore, Harvey argues that firms assert and maintain monopoly rents under the current globalization context through a variety of strategies including concentration, securing property rights and diversifying strategies (“The Art of Rent”).

Our study focuses on the ways in which the documented behaviour by the large publishers studied in this paper coincided with the concentration and diversification strategy suggested for the maintenance of monopoly rent. Concentration of academic content coincided with the centralization of an asset under massive corporatations. Harvey refers to the diversification strategy “as monopoly privileges from one source diminish so we witness a variety of attempts to preserve and assemble them by other means” (“The Art of Rent” 98). The paper argues that the vertical integration of academic content acts as a simultaneous process of aligning open access with the diversification strategy of big publishers. While the movement towards open access reduces the amount of rent available for capture by limiting the revenues associated with each publication, the publisher movement towards infrastructural control valorises their existing owned content, enables the accumulation of additional open data on a massive scale, and furthers the creation of new rent relations through the provision of data and analytics based on their own content and other open content, while simultaneously generating new data through algorithmic means, enabling the appropriation of additional surplus value in a form of value grabbing.

**Methodologies and Materials**

We documented the economic concentration of major academic publishers over the past 20 years until December 2017 using Factset and Capital IQ1 to extract financial data such as Merger and Acquisition (M&A) and revenues. Factset and Capital IQ were

1. Financial data services accessed through the University of Toronto. Factset and CapitalIQ are subscription-based services providing financial statistics, subsidiary structure, and financial news reporting for public companies. The services report all mergers and acquisitions by the firms exceeding a minimum financial threshold and the database includes a brief description of each acquisition.
used due to the information services’ collection of historical mergers and acquisitions information in an easily extractable format. For mergers in 2018, online news articles as well as press releases were consulted. We also collected data on publishers’ current service offerings as presented on their website and on their financial reports.

Extracted data were aggregated into tables sorted by date consisting of over 340 M&A for Elsevier, 80 for Wiley and over 240 for Informa (Taylor & Francis). Due to the involvement of the major academic publishers’ parent companies within a multitude of fields, as well as their potential simultaneous execution of a multitude of business strategies, only those purchases directly relating to higher education or academic knowledge production are included for analysis for the purposes of this article. The items included have been categorized in terms of academic content and academic services/data analytics. The analysis will look at the extent to which publishers have extended beyond the traditional roles of academic publishing into associated processes throughout the knowledge production cycle. This will include a systematic analysis of the various companies/services acquired or launched by these publishers within the academic knowledge production/higher education fields. The analysis will focus on the implications of a vertical integration of scholarly infrastructure by major publishing corporations. At this point, we have not included Springer-Nature and Sage into our analysis because they are private companies and thus data on their M&A are not disclosed on financial databases. We recommend further research on their activities, perhaps using a different methodology.

**Key Findings on Merger & Acquisitions of Academic Publishers**

The figures below present our findings for the M&As of Elsevier, Wiley, and Taylor & Francis. M&A data was systematically disaggregated and coded according to categories such as content, infrastructure and services and non academic acquisitions. A predominant relationship was found between the transition from academic content towards the acquisition of scholarly infrastructure and services.
Figure 1. Elsevier Mergers and Caps Product Launches

Figure 2. Wiley Mergers and Product Launches
Figure 3. Taylor & Francis Mergers and Product Launches

The figures show that all three companies have shown a historical increase in concentration through the merger and acquisition of academic content. The findings corroborate claims of the disproportionate and continued concentration of academic content and its implications to the oligopolistic tendencies in the industry (Larivière et al.). Beyond the expected concentration of academic content, the findings also brought to light an increasing interest in the acquisition of academic services and data analytics primarily since 2007-2008 by Elsevier and Wiley. A scan of press releases over the past four years showed that Taylor and Francis could also be starting to head in a similar direction (Taylor and Francis, “Academic Digital Research Services”).

The remainder of the paper analyzes the potential motivations and implications following the rising trend by large publishers towards the acquisition of services and data analytics as documented by our findings. The subsequent analysis focuses primarily on the experience of RELX and Wiley since their transition towards the acquisition of services is relatively well established and thus facilitates an analysis of potential implications for the entire industry as well as knowledge production at large. We will first show the extent of the transition towards the acquisition of infrastructure before discussing the economic context under which the transition is taking place and the potential motivations and implications of such a transition.
Vertical Integration of Publishers

Our data suggest that the purchases and service launches of Elsevier primarily relate to the academic knowledge and research production process; while Wiley’s purchases and service launches primarily relate to the higher education value chain, a process related to students’ development throughout higher education.

Elsevier

This section expands on the analysis of Elsevier to look into how the acquisition of academic services appears across the academic knowledge and research production life cycle. This analysis requires an understanding of the various stages involved in the knowledge production process. Figure 4 presents a simplified depiction of the various stages involved in the academic knowledge production process. We divided this into three connected sections; the research process, the publishing process, and the research evaluation process. Elsevier academic services were examined individually to see what part of the knowledge production cycle they were targeting. Figure 5 depicts the results. The findings suggest the extent of Elsevier expansion through its acquisitions of various key infrastructural components, a phenomenon that we analyze as a vertical integration of the academic production value chain.
The Academic Knowledge Research Production Process

Figure 4. The Academic Research Knowledge Production Lifecycle

Figure 5. Elsevier Presence Throughout the Research Lifecycle
Elsevier has acquired and launched products that extend its influence and its ownership of the infrastructure to all stages of the academic knowledge and research production process. The various services and their integration will be analyzed in detail in the following section to demonstrate how they influence institutions and individuals’ decision making.

**Wiley**

Wiley’s acquisitions demonstrate a similar strategic move towards a vertical integration for the educational life cycle. To visualize Wiley’s acquisitions of and product launches within the education sector, an understanding of the higher-level education cycle is similarly needed. Figure 6 showcases a simplified depiction of the various stages involved in the university education process. Stages one through seven present a simplified version of the various traditional steps taken by students and designed by academic institutions in the process of obtaining a degree. Simultaneously, important student oriented activities such as student retention, student recruitment and course feedback are also a part of the diagram. Finally, step 8 represents the influence of post secondary education on career development. Wiley services were examined individually to see which sections of the education cycle they were targeting. Figure 7 depicts the results.

**Figure 6. The Education Lifecycle**
As hypothesized, Wiley exhibits a similar vertical integration within the education lifecycle by its acquisitions and product launches. By expanding into additional parts of the value chain such as course design, online platforms, and course testing, Wiley is attempting to attract institutions through the dynamics of convenience and financial benefits (further elaborated in the analysis section). The participation of institutions within the Wiley vertical integration similarly enhances dependency for institutions on the expanded services and products provided by Wiley. Thus, Wiley’s strategy similarly attempts to enhance its position in the education market, especially for online education.

**Analysis**

The vertical integration of scholarly infrastructure has often been justified under the promise that integration allows for efficiency, interoperability and a better-functioning system of services and products (Schonfeld). However, it is precisely because of the power to integrate products across the value chain that this expansion should be critically examined. While we acknowledge the many benefits of a better integration of academic services, we problematize the implications that this integration is occurring under the control of large corporations with tendencies of oligopolistic behaviors. As such, it is crucial to understand the extent to which this integration represents an
increasing influence on decision making power by publishers in regards to the type of research that gets produced, the way in which it is produced, and more importantly who gets to produce it. These questions become particularly relevant given a historical record by large publishers to employ rent seeking strategies such as the construction of the paywall and aggressive concentrations of content with direct implications to access to academic content, while creating exclusions and inequality for resource constrained institutions and individuals.

Elsevier has acquired and launched products that extend its influence and its ownership of the infrastructure to all stages of the academic knowledge production process. This raises an imminent concern of a potential conflict of interest. This is especially true when the largest supplier of academic journals is also in charge of evaluating and validating research quality and impact (e.g., Pure, Plum Analytics, Sci Val), identifying academic experts for potential employers (e.g., Expert Lookup), managing the research networking platforms through which to collaborate (e.g., SSRN, Hivebench, Mendeley), managing the tools through which to find funding (e.g., Plum X, Mendeley, Sci Val), and controlling the platforms through which to analyze and store researchers’ data (e.g., Hivebench, Mendeley) (Elsevier, “All Elsevier Digital Solutions”).

The conflict of interest arises because the influence that Elsevier has over institutional and individual decision making, based on their recommendations and metrics, can privilege their own content as well as researchers that participate within their integrated systems. As such, the conflict of interest has direct implications for the power and control that publishers have over the content and methodological approach of the research being produced. Having publishers involved as key stakeholders of research development and evaluation, as well as researchers’ career development, further raises a number of important questions in relation to the ethics and the negative potential implications of having such a concentrated and centralized force behind the production of academic knowledge.

7. https://www.hivebench.com/
10. https://www.mendeley.com/funding
Specifically, we argue that the vertical integration has the potential to increase dependence by users (universities, lecturers and researchers) on big publishers such as Elsevier and Wiley, as well as transforming the institutional and individual decision-making process, ceding over increasing control to the private for-profit industry. In turn, the increased dependence by individual researchers and institutions as a result of the integration translates into an exacerbation of power and control by the big publishers. Through this logic the rebranding of Elsevier as a data analytics company is also a move towards a future of disproportionate ownership of data by the industry’s giants. A disproportionate ownership of data facilitated by a vertical integration and coupled with an existing disproportionate ownership of content can thus have dire implications for reproducing inequalities across institutions around the world.

The following sections expand upon this arguments by looking at the

- political economic contexts under which the integrations are taking place;
- the dependence generated on academic institutions;
- the dependance generated on individual researchers
- and finally the inequality implications for Global South institutions.

**Political Economic Contexts of Vertical Integration**

While the vertical integration of Elsevier’s products in the academic production cycle brings the potential for greater convenience, the products themselves build a clear dependency through its integration into the institutional workflow. Elsevier’s further move to offering metrics-based decision making, as with SciVal (Elsevier, “See what SciVal can do for you”) and Pure (Elsevier, “Features”), is simultaneously a move to gain further influence in the entirety of the knowledge production process, as well as to further monetize its disproportionate ownership of content, in addition to cementing an increasing supply of content through such dependence. Here, dependency upon Elsevier’s metrics is enhanced as Elsevier leverages the university’s desire to compete with other universities in global university rankings. Indeed bibliometric data from Scopus are heavily used by university ranking companies such as the Times Higher Education whose multiple of league tables are increasingly influential around the world.¹⁴

For universities keen on raising or maintaining their rankings, publishing in Elsevier high impact journals may help them gain the advantage. The promotion of impact

¹⁴. See https://www.elsevier.com/connect/discover-the-data-behind-the-times-higher-education-world-university-rankings
metrics and high impact journals favors established titles within the collection of
titles which Elsevier and the other large publishers have had the time and
capacity to acquire. In fact, the promotion of publication within high impact jour-
nals further entrenches their status. This potential valorisation of existing content
owned by Elsevier is complemented by the capture and creation of new data through
a combination of existing data as well as data extracted through the ownership of
knowledge infrastructure.

The analytics offered by SciVal, such as the visualisation of “topic prominence,” the
fields in science with “the greatest momentum,” serve as a prime example (Else-
vier, “Topic Prominence in Science”). Thus, while additional value may be generated
through the use of algorithms to consolidate certain data, primacy should be placed
upon the ownership of data as it is the core of new value extraction. Thus vertical
integration and the promotion of citation metrics and algorithmic recommendations
may, in fact, constitute rent-seeking behavior designed to increase the dependency
of products, further monetize content ownership, and entrench established journals
within the publishers’ database.

Finally, it is with vertical-integration and the promotion of metrics, as a quantification
of reputation, that publishers are able to co-opt and monetize open access to a grea-
ter extent, with the move towards infrastructure and open access as simultaneous
and complementary processes.

**Dependency at the Institutional Level**

At the institutional level, the vertical integration of infrastructure generates added
dependence on publishers as a result of intense inter-university competition, both at
the local and at the global level. The integration of products particularly in regards to
research evaluation, analysis of the availability of funding, and multiple forms of uni-
versity rankings, create the condition such that universities success and reputation
becomes dependant on their participation within the integrated structures offered by
publishers.

In particular, Elsevier has acquired or developed products such as Bepress (Bepress,
“Features”), Pure (Elsevier, “Features”), Plum (Plum Analytics), and SciVal Funding
(Elsevier, “Research Intelligence”), ostensibly to optimize the university workflow, but
also to capture institutional data throughout the research life cycle. Pure embeds
Elsevier within the university workflow process through its abilities to manage
research at the university level, including the provision of a dashboard to facilitate
decision making by university research administrators (Elsevier, “Features”). Being
able to generate targeted funding recommendations using the Elsevier Fingerprint engine, an algorithmically driven product produced through publication and submission text mining (Elsevier, “Elsevier Fingerprint Engine”), Pure enables algorithmically driven flow of recommendations, and thus enhances Elsevier’s control and influence on institutions’ strategic decisions.

In particular, Pure offers a dashboard as well as other statistics for universities to base their decision making (Elsevier, “Features”), while Plum offers an entire suite of data analytics (Plum Analytics). Bepress promotes the use of “Impact Analytics” for universities to engage in trends and tap into collaborative opportunities in addition to the use of Data Analytics as a method of showcasing the “vitalness” (quote from Bepress) of work (Bepress, “Impact and Analysis”), though vitalness is not defined by Bepress.

Promoting the use of metrics in the decision making process under the promises of efficiency (Elsevier, “Features”), Elsevier further serves to influence university decision making by acting as the creator of source data through its algorithms. Here, in adopting Elsevier products, universities simultaneously cede a portion of their control to Elsevier.

In this sense, SciVal further influences decision making by offering metrics that compare the performance between universities as well as recommending potential collaboration partners at the university and individual level (Elsevier, “See what SciVal Can Do”). Here, Elsevier has provided case studies, including one where an Australian university adopts SciVal to analyse trend data by discipline and conduct scenario modelling for researcher recruitment and retention, with SciVal further “raising awareness” (quote from Elsevier) of opportunities for future collaborations with other institutions (Elsevier, “Case Study”).

The epitome of Elsevier’s attempts to influence university decision making comes through Elsevier Analytical Services, whereby Elsevier’s in-house team provides a report of the university’s performance amongst its peers (Elsevier, “Who Uses Analytical”). Specifically in the offering of “Institutional Research Performance Studies”, Elsevier provides universities the opportunity to hire them as consultants to analyze the university’s performance and makes the corresponding recommendations (Elsevier, “Who Uses Analytical”), cementing their influence in university decision making.

With regard to Wiley, the company enhances dependency by taking advantage of the rise of online degree programs through the offering of an end-to-end online degree consulting and management system (Wiley Education Services, “Higher Education”).
Beginning in the program design stage, where market-oriented strategy is offered, and culminating in education and testing on their in-house Learning Management System (Wiley Education Services, “Higher Education”), Wiley’s attempt is to enhance its control of the online education infrastructure. Here, the competitiveness of a university’s online programs are enhanced, with the expertise/infrastructure of Wiley, while dependency is raised in the use of such expertise/infrastructure. Ultimately, control is ceded in the online education process to the developers of the infrastructures and the recommendations of the experts, Wiley. As such, as institutional attempts to enhance their competitiveness, whether in research or education, institutions become ever more dependent on the infrastructure of Wiley and Elsevier.

Wiley also attempts to influence education strategy directly with its consulting, describing their services as market demand based analysis (Wiley Education Services, “Higher Education”). Here, although Wiley may provide the existing experience in online education, the focus of market demand in program development may lead to a focus on “hot” programs, with Wiley explicitly releasing an article on the viability of online MBA programs (Wiley Education Services, “While On-Ground”). Thus Wiley’s influence extends beyond the integration and alteration of existing programs but rather attempts to encourage an explicit direction in the strategic path of the university, here in the direction of online MBAs for schools with business school programs. There is thus a clear attempt by Wiley to enhance its control over the university decision making process in education, as Elsevier has for academic knowledge production.

With the various product offerings in the Academic Production Cycle, Elsevier enrenches dependence upon its products and enhances its influence in institutions in ways paralleled by Wiley’s activities within the Education Value Chain. Yet despite the parallels in their behavior, the strategic orientations are distinct. In fact, they also represent the need for oligopolistic structures to avoid direct competition. Even as Wiley has moved towards the acquisition of infrastructure situated within the knowledge production cycle (Atypon), their primary acquisitions remain those within the separate education space. Yet the final objective of both companies is the same, being the use of vertical-integration to extract additional value from higher education. With Elsevier in particular, said activity is a form of rent-seeking simultaneously complementing its adoption of Open Access.

**Dependency at the Individual Researcher Level**

At the individual level, the “efficiency” generated through the integrated system also generates a form of network effect. As a result, researchers may perceive
their success as being dependent on participation in this exclusive system. The forward and backward integration with other in-house products makes it so that collaborations, fundings, job prospects and career success are integrated within their use of the entire product suites.

While Elsevier’s products such as Mendeley Reference Manager (Mendeley, “Get On Top”) and Hivebench lab notes (Elsevier, “Features and Benefits”) encourage the use of other integrated Elsevier products, “My Research Dashboard” showcases an individual oriented vertical integration as a free dashboard service for any author who has published in at least one Elsevier journal (Elsevier, “My Research Dashboard”). This personal dashboard “shows download activity, shares, citations, demographic data about who is reading your research, and how and where your publications are being discovered.” For researchers labouring under the “publish or perish” regime, they now have to also demonstrate their citation impact relative to other researchers. This dashboard promises such personalized metrics.

In addition to metrics, publisher could influence individual’s decision with services that curate and promote particular research topics (SSRN), specific opportunities for collaboration (Mendeley, “Join a Community”), and the conducting (Mendeley, “Share and Discover”) and publication (Elsevier, “Find the Perfect Journal”) of their research. The influence of publishers on individual decision making is exacerbated by the aforementioned dependence and influence of services at the institutional level. The dependence to and influence of publishers over institutional strategy has the further potential to permeate into the pressures and incentive systems which are put in place to measure and reward individual success within those institutions. The previous example of an Australian university’s use of SciVal to model researcher recruitment is a case in point (Elsevier, “Case Study”).

SciVal Funding may have great influence on researchers through its ability to recommend targeted funding opportunities as well as potential collaborators to enhance the possibility for maximum funding (Elsevier, “Research Intelligence”). Similarly, Elsevier Journal Finder (Elsevier, “Find the Perfect Journal”), using Elsevier Fingerprint Engine (Elsevier, “Elsevier Fingerprint Engine”), recommends to individual researcher the most suitable journals for research publication, journals with the greatest opportunities for success in terms of acceptance. While services may be used individually, the cross-service integration of Elsevier services encourages the collective use of services, due to additional convenience, thereby maximizing Elsevier’s influence at the individual level.
In fact, the recent acquisition of Aries (Elsevier, “Eslevier to Acquire Aries”) by Elsevier reflects not only an expansion into the space of other publishers but also enables the further acquisition of researchers journal submission data and creates a potential conflict of interest due to the platform’s capacity to redirect failed submissions to collaborator journals (Aries Systems), especially with Elsevier now being simultaneously a publisher and the submission service host for other publishers. In influencing individual researchers in research direction and submissions, Elsevier has the potential to directly enhance the quality of its journals’ content, using data analytics and services to valorise their content while also generating new data through its ownership of the integrated infrastructure.

Individual researchers thus face a double burden of pressures, through the network effect mentioned earlier, but also through institutional incentives, to adopt and rely upon the integrated systems of infrastructures by publishers. A relational power structure is created in which the increasing power of publishers represents a diminishing of choices by individual researchers. This is a distinct disadvantage for researchers in departments or institutions that can not afford to use the integrated system, or who choose to conduct research and collaboration outside of the system.

In parallel, Wiley’s offering of Course Workflow services such as WileyPLUS, provides online testing and is integrated with Wiley’s textbooks through the features of online tutorials (John Wiley & Sons, “Empowering”). It reduces the need for lecturers’ development of questions. Wiley further provides features such as Custom Select (John Wiley & Sons, “Wiley Custom Select”) which simplifies the content selection process within the Wiley content ecosystem.

In Wiley, the convenience of its integrated system may reduce the willingness of lecturers to leave, potentially leading to certain topic compromises. Additionally, individuals are influenced by institutions to alter the structure of courses in incorporate technology such as WileyPLUS due to the potential course savings, with the reduction in Teaching Assistants needed in grading or teaching. Furthermore, End-to-End online degree consulting and management services such as Wiley Education Services (acquired as Deltek) has Wiley consult at the course level (Wiley Education Services, “Higher Education”). Here, Melbourne Business School, the business school of an Australian public research university, serves as an example of the increasing influence of Wiley within the academic education process. Wiley has been hired to:

incorporate digital technology into the learning process... The transformed subjects will have a more individualized, student centric learning process to
improve student learning outcomes and be more efficient for both students and faculty. (Wiley Education Services, “While On-Ground”)

Here, control in the hands of the individual lecturer is further reduced with the top-down instituted Wiley consulting, with Wiley gaining additional influence over the learning process. Ultimately, services provided by Wiley reflect not only an additional source of income but it also has the potential to create environments favouring their other products/content, generating dependencies and potentially enabling further rent extraction.

In both Elsevier and Wiley, institutions and individuals are encouraged to adopt the services due to inter-institution competition. In Elsevier, funding competition and the need for publication and university rankings drives their adoption, while in Wiley, cost reductions and competition between institutions for students serve as the rationale for their adoption. Yet in both cases, control is transferred to the publisher as their recommendations/consulting become increasingly “crucial” to achieving the goals of institutions/individuals’ success, whether funding/publication or enhanced enrolment in education; with the motivations of the institutions in particular tied into the global university rankings.

At the same time the integration of services by large publishers and the resulting dependence makes it harder for alternative services to emerge or succeed, particularly since they lack the existing system of support and compatibility. In addition, alternative services have to compete with the substantial financial cross-subsidies and the disproportionate ownership of content and data that big publishers possess.

**Exclusionary Implications towards Global South Journals and Researchers**

The dependence and increased reliance on publishers’ products, services and infrastructure give publishers a higher stake at influencing how institutions value and incentivize research and education, with direct implications on the direction of research investment and resource allocation. This influence is intensified by the aforementioned disproportionate ownership of data collecting infrastructure, which gives publishers the capacity to generate algorithmic recommendations through which to enhance their control and influence over the institutions.

As publishers become key influencers in terms of how research gets produced and who gets to produce them, the implications for inequality within the university and globally can be expected to rise. Institutions that can not afford or are not interested
in using the products and services offered by the publishers face a vast disadvantage with regards to their competitiveness in attracting future funding. Additionally, in their attempt to achieve success, such institutions may end up replicating and following publishers’ defined models of success. Inequality would be particularly worsened for resource constrained institutions in the Global South, as the narrow measure of success (for example publication output, citation impact, and university rankings) promoted by publishers becomes the universal yardstick for all institutions.

The dependency on the products and services of publishers such as Elsevier as well as the promotion of citations metrics have the potential of furthering inequalities and marginalization of global south journals and researchers. This is because products such as SciVal, Mendeley, and SSRN promote collaboration with their potential paying customers, who are primarily based in well resourced institutions in the Global North. The matching features of the Elsevier Fingerprint Engine is also limited to English, further excluding non-English speaking Global South researchers from a technology that Elsevier is increasingly promoting.

Furthermore, the funding targeting abilities and comprehensive databases within products such as SciVal Funding can possibly increase the success of Global North researchers while excluding researchers from the Global South. It is also well known that the promotion of citation metrics favours Global North journals, who have benefited from the western modality of research as well as the western norms of scholarship, to the potential exclusion of global south scholarship (Chan and Costa; Tijssen et al.; Nwagwu). For Global South scholarship to join the discourse and to participate, there exists an increasing need to adopt the western norms of scholarship, and an increasing allure for Global South journals in joining a Global North publisher and be included in the primary indexing databases such as the Web of Science and Scopus (Beigel; Ráfols et al.; Le Roux). Joining a Global North publisher, some have argued, serves as a form of academic neo-colonialism, as the Global North firm will exert direct influence upon the policies of such journals; while the adoption of western forms of scholarship merely enhances the hegemony of Global North academia (Okune et al.; Fiormonte and Priego). Both processes thus add new layers of marginality to Global South epistemologies (Nyamnjoh; Hall and Tandon). Such is only a very precursory discussion of the exclusionary implications of the increasing concentration of control of infrastructure by dominant publishers, and we continue to explore these issues with the Knowledge G.A.P. project.  

---

15. See http://knowledgegap.org
Conclusions

Having documented the historical increase in concentration within the academic publishing industry through mergers and acquisitions, we further described the expansion of publishers towards associated academic services and data analytics by building end-to-end infrastructure that span the entire knowledge production life cycle. At the same time, they are embracing open access, and their clear intent is to leverage their disproportionate content/data ownership to enter into a new field of data analytics with existing competition, most notably Clarivate, who owns the Web of Science.

We have shown the strategic intent of both RELX and Wiley in diversification. The differing primary expansion paths for the two companies may be a form of oligopolistic behavior to not only leverage their possessions but also to enhance monopoly power within each lifecycle, whether in education or knowledge production.

Thus despite the superficial benefits of the rising uptake of certain open access practices by big publishers, it is clear that they are developing complementary methods of entrenching dependency by institutions and individual researchers. We expect the exclusionary forces generated by the increasing adoption of Elsevier products to have a direct impact on the diversity of knowledge production, with exclusion on funding, citations, norms of discourse, and research collaboration. This is a key area and a recommended direction for further research.

Vertical integration has the further potential to exert a direct exclusionary effect on less financially well-endowed journals and institutions, primarily those in the Global South, in their attempt to emulate the western modality of knowledge production. To address global inequalities in knowledge production and scholarly representation there is a clear need for a community-driven integration of scholarly infrastructure, one that is designed to mitigate inequality, that serves the public good within the community, rather than one which only seeks to maximize profits and co-opt open access for the objective of rent-seeking.

References

corporate/elsevier-to-acquire-aries-systems-a-best-in-class-publication-workflow-solutions-provider
Fuller, Steve. The Academic Caesars: University Leadership is Hard, Sage, 2016.


The End of a Centralized Open Access Project and the Beginning of a Community-Based Sustainable Infrastructure for Latin America

Redalyc.org after Fifteen Years

Arianna Becerril-García
Autonomous University of the State of Mexico
arianna.becerril@redalyc.org

Eduardo Aguado-López
Autonomous University of the State of Mexico
eduardo.aguado@redalyc.org

Abstract: The Latin American region has an ecosystem where the nature of publication is conceived as the act of making public, of sharing, not as the publishing industry. International, national and institutional contexts have led to redefine a project—Redalyc.org—that began in 2003 and that has already fulfilled its original mission: give visibility to knowledge coming from Latin America and promote qualitative scientific journals. Nevertheless, it has to be transformed from a Latin American platform based in Mexico into a community-based regional infrastructure that continues assessing journals’ quality and providing access to full-text, thus allowing visibility for journals and free access to knowledge. It is a framework that generates technology in favor of the empowerment and professionalization of journal editors, making sustainable the editorial task in open access so that Redalyc may sustain itself collectively. This work describes Redalyc’s first model, presents the problematic in process and the new business model Redalyc is designing and adopting to operate.

Keywords: Redalyc, open access, Latin America.
The Open Access Ecosystem in Latin America

Although Open Access (OA) is a worldwide reality, it displays various levels of growth in each region, leading to uneven regional impacts.

Latin America, for example, has created and maintains a noncommercial structure where publishing belongs to the academy and not to large publishers, where the Open Journal Systems software has been key in the birth of the electronic journal, where the need for visibility, interoperability and presence on the web was the breeding ground for the emergence of platforms such as Latindex, Redalyc and Scielo. The Latin American region, as a result, owns an ecosystem characterized by the fact that “publishing” is conceived as acts of “making public”, of “sharing”, rather than the activity of a profit-driven publishing industry.

A fee has not been included neither for authors nor for readers in the regional editorial tradition. Fees where they may exist are definitely designed outside of a profit motive, the reason being that the corresponding journal business model usually relies on institutional funds.

Latin American academic journals are led, owned and financed by academic institutions. It is uncommon to outsource editorial processes.

In Latin America, each academic institution, de facto, is part of an informal cooperative system that is neither formalized nor made explicit. Every institution supports journals that are driven by their own faculty members, and then that content is made available to other institutions in OA. Which means that everyone benefits from everyone’s investment. This kind of informal cooperative was already working before Open Access got its present label.

Regarding Green Open Access, the region has different initiatives such as the growing quantity of institutional repositories that currently reaches 362 (OpenDOAR). For its part, LaReferencia, the network of national networks of institutional repositories, integrates through metadata harvesting more than 1,3 million documents.

In the Latin American region, a number of laws in favor of Open Access have been voted in several countries, including Peru, Argentina and Mexico. According to Roar-Map, 48 mandates are available at national and institutional level (ROARMAP). However, as it is pointed out by Babini institutional mandates tend to be weak.
Disciplinary repositories have also emerged, such as CLACSO, specialized in Social Sciences, SIDALC in Agriculture, CLAD in Public Administration, LaborDoc in Labor, BVSDE in Environmental Health and BVS in Health.

![SWOT Analysis](Figure 1. SWOT Analysis. The Open Access ecosystem in Latin America)
From our perspective, there are many strengths but also important threats that are shaping different aspects of how OA is evolving in Latin America. The illustration below shows a SWOT analysis of the OA landscape in Latin America.

A crucial weakness is that the majority of research assessment systems rely on simplistic metrics which, ultimately, rest on the position of journals in the JCR rankings of Clarivate Analytics or Elsevier Scopus, regardless of the scientific conversation taking place within the journal and of the local or regional problems being solved or discussed. Researchers are strongly impacted since they are judged by the journals they publish in, rather than what they publish. Colombia offers a good example of this flawed approach: in 2017, it eliminated 40% of its national publications from its national index, Publindex.

Most Colombian, as well as Latin American journals, are published in universities from the educational sector and as a consequence of such decisions they are receiving less and less resources each day, causing thus a devastating disappearance of journals which are not indexed by JCR or Scopus, ignoring the journal’s contribution to science and society.

The same happens with Redalyc, as it is an index that is out of what is considered the status quo in research assessment systems, its financial supporters start to question the feasibility to continue investing in this project overlooking the ten million article downloads per month or the 54,000 users accessing to Redalyc per day.

This problem, along with the lack of editors’ leadership and the researchers’ carefree attitude with minimal response to research assessment policies, puts at risk the legitimation of both regional and Open Access initiatives that could function as a counterweight to commercial publishing.

Scheliga and Friesike analyze the fact that while many researchers support Open Science in theory, the individual researcher is confronted with various difficulties when putting Open Science into practice. They argue that the phenomenon of Open Science can be seen through the prism of a social dilemma: what is in the collective best interest of the scientific community is not necessarily in the best interest of the individual scientist.

As it is said in DORA:

The outputs from scientific research are many and varied, including: research articles reporting new knowledge, data, reagents, and software; intellectual
property; and highly trained young scientists. Funding agencies, institutions that employ scientists, and scientists themselves, all have a desire, and need, to assess the quality and impact of scientific outputs. It is thus imperative that scientific output is measured accurately and evaluated wisely.

Another circumstance to analyze is the lack of innovation, professionalization and development of technology in the regional institutional publishing. The big OA platforms are given the role of providing publishers with technology and innovations.

Talking about big regional OA platforms, currently each one is going in a different direction, preventing a strategic regional coalition to be possible. Such is the case of Scielo, which opted to look towards Clarivate Analytics in order to create a journal citation index inside Web of Science (out of the core collection), a strategy to achieve the inclusion of journals in the mainstream science. On the contrary, Redalyc seeks to strengthen publishers inside universities by empowering editors, with technology and training, and providing a set of metrics that show different aspects of the research performance rather than impact based on citations.

Redalyc’s vision is more aligned with Curry who argues that it is time for academics to take back control of research journals and he underscores that the evolution into a highly-profitable industry was never planned. Academics must make the case for lower-cost journals.

Apart from that, there are some “external” phenomena affecting the noncommercial nature of the Latin American ecosystem. After the formalization of the Open Access movement through the three statements: Budapest (Chan et al.), Berlin (Max Planck Society), and Bethesda (Brown et al.); the Global North started to transform its editorial practices, which focus mainly on commercial publishing in order to achieve OA. Those changes caused the relocation of costs, from readers to authors with the so called Article Processing Charges or APCs. The big four players in the publishing industry: Elsevier, Springer, Wiley and Taylor & Francis, have all embraced OA, to varying degrees. They have also employed highly different strategies as to how much they would like to charge their authors (Socha).

Since then, researchers have witnessed an increase in APCs, a circumstance that hinders Latin-American authors to publish. Besides, as it was said by Katie Shamash, the number of article processing charges paid doubled between 2013 and 2014. The average APC increased by 6% in those years, a rise well above the cost of inflation. Publishers’ APC costs are converging to a more uniform price range, although they
still vary widely. Journals with low APCs are raising their prices, perhaps to avoid being perceived as low quality following expectations set by Finch’s report.

As asymmetries between the for-profit and non-profit OA are extended, corporations impose rules that restrict access to knowledge in multiple ways. Gadd and Covey’s paper demonstrated this by analyzing the Sherpa/ROMEO database:

> Just as there is an upward trend among these 107 publishers in the number of publishers allowing some form of self-archiving, there is also a year-on-year rise in the number of restrictions and conditions constraining the right to self-archive and the offer of paid open access options. . . Restrictions around when a paper may be self-archived grew 1000% . . . Restrictions relating to where a paper may be archived were even more prevalent and followed a similar growth pattern of 190%.

This work describes Redalyc’s first model that led the project to achieve its original goals, presents the problematic it is facing and the new business model Redalyc is designing and adopting to operate. A new configuration of strategies, in response to the international, regional, national and institutional contexts, that seeks a collaborative, sustainable, protected and noncommercial Open Access solution for Latin America.

**Redalyc as a Centralized OA Project**

Redalyc began in 2003 with the main goal of contributing to the visibility of scientific journals published in the region, in a time where the majority of them did not even appear on the Web. It started indexing only journals of Social Sciences and Humanities and in 2006 included all areas of knowledge owing to a high demand from editors.

Today, Redalyc’s collection contains more than half a million full-text articles from 1,260 Open Access peer-reviewed journals published by 622 publishers from 22 countries of Latin America, Spain and Portugal.

The following diagram shows Redalyc’s infrastructure. Each layer presents different added-value services that Redalyc provides to journals so as to complement the features and capabilities the institutional publishers are able to achieve. It is important to highlight that in other regions of the world there are commercial publishers in charge of providing these kind of services.
On the basis of the structure shown there are peer-reviewed journals along with their publisher institutions who sign an agreement with Redalyc in order to allow electronic distribution of metadata and the online publication of full-text articles.

For a journal to be indexed in Redalyc it must pass a rigorous selection process based on quality criteria (Redalyc, “Criterios”) including among other mandatory requirements: to carry out a peer-reviewed process and to publish at least 75% of original content, joint with criteria such as the variety of referees and editorial board, the efficiency of editorial processes, editorial, visibility and technological practices and periodicity compliance. After getting an internal result, the International Advisory Board, which is composed of experts from diverse fields and different countries, is asked for a qualitative assessment.

On the other hand, Redalyc is aware that one of the highest costs in publishing electronic journals is the XML tagging and that this process is key in reaching technological standards. Indeed, Redalyc developed an XML markup system—called Marcalyc—in order to contribute to the sustainability of journals (Redalyc, “Marcalyc”).

Marcalyc is based on the Journal Article Tag Suite ANSI/NISO Z39.96-2015 standard (NISO) and allows journal editors to get its articles in XML file format. Free access to this tool is provided for Open Access non-APC scholarly journals indexed by Redalyc. It is a tool designed to prevent editors from outsourcing XML markup; it does not require technical expertise and it minimizes markup time. It is also compliant with JATS4R (JATS4R).
Marcalyc, together with the XML file format, automatically generates a media enriched article reader and a mobile reader available in Redalyc as well as the PDF, ePUB and HTML versions ready to be uploaded in journal websites.

Since Marcalyc was launched in September 2016, 1,158 journal issues have been processed. In a collaborative process, Redalyc provides the tool and the journal editor does the job. Based on the average market cost of XML generation per journal issue (US$300), that quantity of XML files would have cost approximately USD$347,400. For each of all those articles, Redalyc also provided the PDF, ePUB, HTML and the interactive article reader versions, which costs must be summed to the total subsidy that Redalyc is providing to journals in Latin America.

Coupled with that, there are successful cases that show Redalyc’s contribution, e.g., journal editors generating XML content with no-cost in Marcalyc, and taking them to their own websites, along with all file versions listed above; and journal editors switching from a policy of APC to a non-APC in order to apply for a Marcalyc user account.

What is more, Redalyc provides homepages not only for journals but also for countries, authors, institutions and areas of knowledge with their journals collection, search engines, advanced filters, different data visualizations and metrics.

Regarding visibility, Redalyc works with search engines, libraries, directories and social media to disseminate broadly the scientific content and to improve its discoverability. Redalyc also works with repositories and aggregators through interoperability protocols such as OAI-PMH.

Besides, the signed agreements between Redalyc and journals enable the former to populate institutional repositories too, helping repositories to automatically integrate metadata of articles corresponding to the institutional output. For example, in the case of the National Autonomous University of Mexico (UNAM), Redalyc is able to send more than 10 thousand articles at metadata level to its institutional repository.

In different dimensions that is the case of more than 10 thousand institutions with research output published in journals indexed by Redalyc: they can take advantage of Redalyc’s database to populate their repositories.

The same happens with the scientific output at a country level as shown below, data which may be used to strengthen national repositories.
Redalyc has also developed a full integration with ORCID. In the ORCID Registry, authors can use a wizard created by Redalyc to connect their ORCID ID to works published in journals indexed by Redalyc. In addition, the platform called RedalycAuthors enables researchers to create a scholar homepage, monitor article metrics and export works to an ORCID profile.

Redalyc is one of the thirteen databases that have done a full integration with ORCID, including ResearcherID from Clarivate Analytics and Scopus from Elsevier. This effort contributes to the insertion of Latin American authors in the global scientific conversation.

Finally, Redalyc offers an alternative set of metrics, which are intended to answer issues, such as institutional strengths, influential journals in a knowledge area, use of the knowledge generated by an institution, cohesion among institutions in terms of collaboration and trends in the internationalization of scientific output, among others.

The Study of the Use, Assessment and Impact of the National Consortium of Scientific and Technological Information Resources of Mexico (Scimago) shows that Redalyc’s website is one of the most used sources of information in Mexico, preceded by CONRICyT (National Consortium of Scientific and Technological Information Resources)—which groups all subscription databases Mexico is paying access for—, Scielo Mexico and the journals’ websites, even though Redalyc does not receive any national financial support. Indeed, CONRICyT and Scielo Mexico are sustained by the National Council of Science and Technology CONACyT.
With all this growth appears an urgent need to extend Redalyc’s institutional support to other stakeholders rather than rely only on the Autonomous University of the State of Mexico, who has supported Redalyc since its beginning, yet it is aware that Redalyc’s current reach exceeds the university’s capacity. It is important to recognize that a project with that regional scope must not depend on a single actor in order to guarantee long-term sustainability.

These circumstances have led to redefine a project—Redalyc.org—that begun in 2003 and which has already fulfilled its original mission: give visibility to knowledge generated in Latin America and promote the quality of scholarly journals. Nevertheless, it is mandatory to be transformed from a Latin American platform based in Mexico into a community-based regional infrastructure that continues assessing journals’ quality and providing access to full-text articles but taking into consideration individual journals’ efforts for the sake of journals’ positioning and visibility. It is a framework that generates technology in favor of the empowerment and professionalization of journal editors, making the editorial task in open access sustainable and which allows Redalyc to sustain itself collectively.

The Next Redalyc: A Community-Based Sustainable Infrastructure for Latin America

It is necessary not only to refine critical positions to face the global context of OA, but to design and put into practice collaborative systems that develop and socialize technology and know-how in scientific communication.

The goal is to build and sustain a collaborative infrastructure to develop technology and generate knowledge in order to strengthen and protect the scientific editorial work and its actors, a framework that can be extended to include the Global South, with a special emphasis on social sciences and humanities. All this to achieve a non-subordinate integration of this region into the global dialogue of scientific communication.

This is aligned with Johnson’ vision, where it is said that:

. . . we should work towards simplifying and standardizing processes to move towards a sustainable and scalable OA ecosystem which preserves academic freedom and author choice in publishing and makes the research as valuable as possible for the end user.
The new strategy consists of forming an alliance of institutions that are willing to maintain the noncommercial nature of the Latin American ecosystem and keep supporting scientific publications that contribute to benefit society and the development of science, regardless of whether they are indexed in the so-called mainstream science or which impact factor is assigned to it.

The participating institutions should share the following DORA declaration themes (DORA, 2012):

- the need to eliminate the use of journal-based metrics, such as Journal Impact Factors, in funding, appointment, and promotion considerations;
- the need to assess research on its own merits rather than on the basis of the journal in which the research is published;
- and the need to capitalize on the opportunities provided by online publication (such as relaxing unnecessary limits on the number of words, figures, and references in articles, and exploring new indicators of significance and impact).

In order to achieve this transformation, the most relevant tasks that have been established are:

- recognize Redalyc as an inclusive indexing journal system that contributes to the strengthening of scientific journals’ quality published in the region;
- develop and adopt a set of responsible metrics for the assessment of scientific performance that takes into account the contribution made to the field of knowledge and society. This task includes the selection of the most suitable metrics already developed by stakeholders, or perhaps, develop new ones;
- develop and sustain a digital edition tool that enables collaborative paper XML tagging under JATS metadata standard that provides regional publishers with the possibility of being at the forefront of technology at minimal cost;
- develop technology and provide knowledge transference for tasks such as digital preservation of scientific content, visibility, interoperability and discoverability, among others;
- develop technology for the optimization of editorial processes through the promotion of forming a developers’ community in order to take better advantage of the OJS;
- build a blog as a means for critical discussion on scientific communication and its particular problems in the Global South: challenges, actors, trends, news, advances, reviews and proposals.
The institutions that up to now have expressed their interest in taking part in this model are University of La Plata, Argentina, University of Antioquia, Colombia and CLACSO, Argentina, among others.

Some of these institutions also intend to contribute with computing power, software developers and high-speed connectivity to enhance the service for users and continue developing innovative solutions.

The new constitution Redalyc will adopt is based on these fundamental principles:

- Scientific knowledge generated with public funds is a common good and access to it is a right. Therefore, it must be available in Open Access. We recognize that Open Access has been used as a tool to benefit large publishing consortiums, it is an unwanted effect of the movement that must be controlled.

- Open Access cannot be consolidated unless it is legally protected. The Open Access mandates and laws, and the use of Creative Commons BY-NC (Non-Commercial Attribution) licenses and desirable SA (Share Alike) are fundamental protection strategies, which are aimed at not allowing the appropriation of scientific knowledge for profit.

- Open Access has no future or meaning without an evolution of research evaluation systems: if we recognize the value of Open Access as an appropriate channel of communication, it must be recognized as a legitimate way from which to evaluate scientific and scholarly practices.

- The consolidation of Open Access must consider the transition to digital scientific communication as an essential axis as well as the development and adoption of technologies that allow to think about new communication possibilities.

- The economical investment in Open Access must be coherent with its benefit to society just as commercial solutions are paid.

- The adverse economic scenarios Open Access faces in Latin America will have to be overcome with work schemes based on collaboration and sustainability; likewise, through a permanent disposition to dialogue with science and technology councils, governments, universities, international organizations and all actors with a shared vision of this movement.

- It is necessary to recognize the diversity of scientific journals and stop the pressures that seek to homogenize them, from language to research agendas. It is necessary to vindicate the relevance of local, or national and regional issues in journals; as well as the articulation with the needs of society’s development and quality of life. Likewise, it is necessary that journals support the strengthening of
institutional repositories by means of removing embargo policies of scientific articles or retention of copyright that prevent the deposit.

- The sustainability of Open Access is the conviction that science can and should have a central role in the development of societies. The social impact of science is the foundation of its existence.

- Different areas of knowledge have different dynamics of generation and circulation of knowledge. It is necessary to respect the different idiosyncrasies by area, especially the dynamics of Social Sciences and Humanities, as well as their different communication formats.

- Open Access must be permanently conceptualized and defined according to the contexts it faces; this is an imperative that arises when observing that the context in which the statements of Budapest, Bethesda and Berlin emerged is different from the current one. The three "B" homogenize the conditions of development of science and the conditions of the South are different from those of the North.

Redalyc will now participate in a regional, multi-institutional infrastructure based on cooperation that allows obtaining resources. This new cooperative infrastructure is AmeliCA (Open Knowledge for Latin America and the Global South), whose main objective is to preserve the open nature of Latin American science communication environment.

With this, Redalyc's technology will be extended to the Global South by opening—in a distributed architecture—the technological developments in cooperation with other institutions, including: the journals markup system, bibliometrics, author pages and enriched reading formats.

This initiative will be composed of a Council, an Executive Secretary and Commissions.

Each multi-institutional commission will coordinate a line of work. Initially, the following commissions are proposed:

- Technological development;
- Research;
- Editorial processes;
- Continuous training;
- Responsible metrics;
- Observatory of scientific evaluation;
Conclusions

It is necessary to think of Open Access from the Global South perspective, recognizing its experience and leadership in the defense and contribution to the movement. By the same token, it is necessary to think of the Global South not only as a geographical region, but also as a shared historical past and as a possibility of future construction. Thus, the defense of Open Access will also be a defense for the memory, justice, truth and development of people.

The future of academic publishing in our region must be a commitment by its actors, as well as a collective right that must be actively and creatively claimed.

References


Gadd, Elizabeth and Denise Troll Covey. “What Does ‘Green’ Open Access Mean? Tracking Twelve Years of Changes to Journal Publisher Self-Archiving Policies.”
doi:10.1177/0961000616657406
Global Scholarly Collaboration
From Traditional Citation Practice to Direct Communication

Sergey Parinov
CEMI RAS, RANEPA
sparinov@gmail.com

Victoria Antonova
Higher School of Economics
v_antonova@mail.ru

Abstract: The development of recent research information systems allows a transformation of citations in the full text of research papers into interactive elements. Such interactivity in some cases works as an instrument of direct scholarly communications between citing and cited authors. We discuss this challenge for research e-infrastructure development including opportunities for improvements in research cooperation and in collaboration mechanisms for the global research community.

Keywords: citations as interactive elements, scholarly communication, cooperation, collaboration, shared mental models.

Introduction

Current achievements of scholarly digital publishing and citation content analysis create new ways for the better support of the Knowledge Commons. In this paper we discuss the integration of some tools from two ongoing projects. One offers a method of transforming citation data into “interactive elements”, another promises to transform the interactive elements into a communication instrument between citing and cited authors.

Both projects are based on the principles of the Open Scholarly Infrastructure (Neylon et al.). This guarantees that even after completing the projects, their outputs will be available as a part of a sustainable infrastructure. It means that if there is a demand for the projects’ outputs, they can be freely re-used and developed by anyone.
To discuss how current scholarly communication, based on publishing and citing papers, can be replaced by a collaboration between researchers based on direct communication, we should start with a clarification of the relationships between scholarly communication, cooperation, and collaboration.

The main tool of scholarly communication for researchers with the research community is that of publishing their research. References in such publications serve as observable evidence of scholarly cooperation in the research community. Scholarly cooperation, as a kind of socio-economic cooperation (Smith et al.; Axelrod and Hamilton) means that some scholars use the outputs of other scholars and by this, they realize the collective development of scientific knowledge.

Citation network analysis allows us to determine whose results the scholar used, i.e., with whom he had a cooperation and the position of the scholar in this network. For the research community as a whole, there is no other common method to indicate what and how outputs were used by particular scholars in their research.

The forms of cooperation used to organize collective activity may vary depending on the number of participants in this activity and other conditions (Smith et al.). Processes of scholarly cooperation based on publishing and citing papers have the form of “horizontal cooperation” (Smith et al.) among the research community.

Scholarly cooperation on a community-wide scale is observable and measurable because of:

- current rules and regulations, which oblige scholars to specify their use of previously published research outputs in a prescribed manner;
- tools and services for research papers preparation and publishing;
- means for processing the full text of publications and visualizing citation data, relationships and statistics.

As a result, the research community has technology to produce and share publications with citations as an instrument of cooperation. There are also formal/informal rules and social institutions that regulate the application of this cooperation instrument.

Let us compare: scholarly cooperation based on the exchange of publications and citations with another well-known form of collective activity based on direct communication between members of a small group like a research laboratory or a project team.
In the first case, the “cooperation is accomplished by the division of labor among participants as an activity where each person is responsible for solving a portion of the problem” (Power). If we look at how a small group works, as in the second case, we see a collaboration, which “is a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem” (Power).

Such comparison shows that publications as a communication and cooperation tool offer participants significantly less effective cooperation. In particular, the authors of research publications do not know about all the facts when someone uses their published research. As a rule, they do not have accurate and complete information on who and when used (cited) their publications, what exactly was used (cited) and for what purposes. There are no common ways for authors to publicly respond to how other researchers used (cited) their publications. The research community does not have a complete picture of how the publication was used to create new scientific knowledge. Similarly for the authors: it is not known exactly which of their outputs were used, by whom and how exactly.

If we consider citations as a reflection of scholarly cooperative links, then in real academic life the citations bear a lot of information, not all of the same weight. These aspects, including scholars’ lines of behavior associated with citations, and the actual content of citations are studied by the sociology of citations (Adler et al.). As the main type, the “grateful” citations (Adler et al.) are pointed out, considered as an acknowledgement of the intellectual debt in relation to the cited publication. Such citations directly indicate the scholarly cooperation between citing and cited authors.

Some researchers, however, consider that for modern research publications the “rhetorical” citations (Adler et al.) are more common (Cozzens). They work as the means of conducting scientific discussion, serve as illustrations and also perform certain “ritual” functions, not directly related to the process of collective creation of scientific knowledge and scholarly cooperation.

This situation is further aggravated by the use of citation indexes to assess research performance. Measuring the success of scholars by the number of citations their publications receive affects the essence of the citation process, because in many cases this indicator becomes the goal of scientific activity. Consequently, all the characteristics of the citation process cease to be trustworthy (Neylon).

Despite this, the publication of research results and the practice of citation are mandatory in the research community. This allows us to suggest that the data on citation
networks extracted from research publications carry information about a substantial part of the existing global network of scholarly cooperation. The processing of these data makes it possible to visualize this network and presumably identify potentially collaborative scholars.

Identifying pairs of scholars, citing and cited, one can suppose that they are related by potential scholarly cooperation. If the contact information of both is available, it is possible to organize actual (direct) scholarly communication between them. When such communication becomes possible between all pairs of cooperative scholars, it promises them better research performance.

The development of a technology for scholarly cooperation between citing and cited authors creates the conditions when researchers can identify the “suppliers” of the research results required by them, and the “consumers” of their own results. Hence, scholars are able to simultaneously coordinate their research in two directions with the “suppliers” of the research results required by them, and with the “consumers” of their own research.

The next step is the creation of the conditions for the emergence of full-fledged collaboration between citing and cited scholars, which normally arises only in small groups.

The second section of the paper considers some research information systems (RIS), the combination of which allows the citations found in research publications to be turned into interactive elements. Such innovation, in its turn, allows RIS to initiate direct communication between cited and citing authors.

In the third section, we discuss how the small-group mechanism of collaboration works and what RIS functionality is required in order to create conditions for the operation of such mechanism on the scale of the entire research community.

In conclusion, the possible consequences of cooperative mechanism implementation for the scientific community as a whole are briefly discussed.

**From Citing of Papers to Direct Scholarly Communication**

There is a clear motivation to create a technology which allows direct scholarly communication between researchers who currently can only cite each other’s papers.
As expected, such an opportunity for scholarly cooperation can greatly increase the research performance of these scholars.

This technology is a combination of existing and newly created tools. Some new tools allow parsing of extended set of citation data from research papers, which also gives us the citation contents around the in-text citations. Within a RIS, the citation content can be used to initiate direct scholarly communication between citing and cited authors. A citation content analysis also helps to use such communications to develop research cooperation.

**Parsing of Citation Data from Research Papers**

The project CyrCitEc¹ is developing a tool for the parsing of citation data, including a context for in-text citation from full-text research papers in PDF (Parinov, “Semantic Attributes for Citation Relationships”). The project is funded by the Russian Presidential Academy of National Economy and Public Administration (RANEPA²). It takes input data from RIS RePEc³ and Socionet.⁴

In late March 2018, CyrCitEc processed 203 collections of papers with 89,342 publications in total. The largest part of this set are 157 Russian academic journals covering different academic disciplines and provided by NEICON.⁵ There are also research paper series in Russian and English provided by Russian universities, including the Higher School of Economics⁶ and RANEPA⁷.

The approach used by CyrCitEc for citation data parsing was presented in “Semantic Attributes for Citation Relationships: Creation and Visualization” (Parinov). All citation data extracted by the CyrCitEc project are publicly available at http://cirtec.ranepa.ru. Regularly updated CyrCitEc statistics about parsing results are available at http://cirtec.ranepa.ru/stats.html.

Only 69% of the papers have full text PDFs available for the citation data parsing and only 51% of the papers have a list of references in a more or less standard form.

---

1. https://github.com/citeccyr
Based on the set of papers with references, we parsed in total 801,318 references, that is on average 18 references per paper. In this set, about 5% of references are duplicated, because different papers can cite the same publications and have the same references. Such low duplication rate is a consequence of a very multidisciplinary character of a processed collection of academic journals.

For 26,467 of the parsed references, we were able to create citation relationships between citing and cited papers, since we found cited papers’ metadata within RePEc and Socionet information systems. It is only about 3% of total references because currently these research information systems don’t represent the majority of research papers.

Using the same set of papers and approach, we parsed 750,607 in-text citations. They mention 1,072,175 parsed references. This is 270,857 references more than the total number of parsed references, since some references are mentioned more than once. On average, there are 1.3 mentions per reference.

Non-mentioned references were also counted: 110,340 references (14%) have no in-text citations at all. About 37% of papers with references have at least one non-mentioned reference.

**In-text Citations**

Among different types of citation data available in research papers, in-text citations may keep the data about the character of cooperation between citing and cited authors.

The in-text citation data parsed by CytEcCyr include the following attributes (see also an example of data record below):

- a text string of the style of in-text citation, e.g., a number or an author name in square or round brackets (the tag `<Exact>` in the example below);
- a link to a reference, mentioned in this in-text citation (the tag `<Reference>` below);
- text coordinates of the in-text citation, i.e., a serial number of the first and the last in-text citation symbols counting from the beginning of the paper’s content (tags `<Start>` and `<End>`);
- citation contexts located on the left and on the right of the in-text citation; these include at least 200 symbols expanded for taking a whole sentence (tags `<Prefix>` and `<Suffix>`).
An example of parsed data about one in-text citation

```
<intextref>
  <Prefix>... countries and Soviet republics</Prefix>
  <Suffix>; Gokhberg, Kuznetsova, 2011]. ...
  <Start>8757</Start>
  <End>8781</End>
  <Exact>Gokhberg et al., 2009</Exact>
  <Reference>20</Reference>
</intextref>
```

The in-text citation from the example above has a link with a reference having the number 20 in a paper. CyrCitEc parsed for this reference the following data:

```
<reference num="20" start="54464" end="54654"
  author="Gokhberg Kuznetsova Zaichenko" title="Towards ..."
  year="2009"
  <from_pdf>Gokhberg L., Kuznetsova T., Zaichenko S. (2009)
  Towards a New Role of Universities in Russia: Prospects and
</reference>
```

All CyrCitEc data with in-text citations and related references are input for the next tool, which is a part of the Socionet RIS.

**Socionet Tools**

Socionet services, as described in the article “Semantic Attributes for Citation Relationships: Creation and Visualization”, use in-text citations and references data to produce computer-generated annotations in the content of PDF papers (Parinov). Figure 1 shows what these annotations look like using in-text citation and reference data from the examples above.
Readers of PDF papers see the in-text citations, if they exist, as annotated text. At figure 1 there are mentions of two references in brackets. These highlighted in-text citations work as interactive elements, since clicking on them opens an information box (below on figure 1) with additional data about the cited paper. The additional data can include details about the cited paper (citing statistics, title, authors, etc.) and links to additional tools.

One of the Socionet features is the multiple semantic relationships between information entities (Parinov, “Towards a Semantic Segment of a Research e-Infrastructure”). It allows the linking of citation data with other types of information. A fragment of the semantic linkage network is presented at figure 2.

Using these linkages, we can associate additional data with the citation data, e.g., the contact and affiliation data of the authors for the cited and citing papers, metadata of cited and citing papers, etc.

As we expect, in 2019 CyrCitEc will process all available publications from RePEc and Socionet. As a result, the parsed citation data will be automatically linked with about 70,000 author profiles. These author profiles are linked with papers, and with about 15,000 profiles of organizations that also have links with other author profiles from their staff.

---

The system, using the linkages, can also recognize types of relationships between signed-in users and information objects:

- a user is the author of the browsed object (paper, citation, etc.);
- a user is a co-author of the selected author;
- a user is a citing or cited author;
- a user is just a reader; they have no relation with the browsed object.

This information allows the system to initiate different scenarios of direct communications (Kogalovsky and Parinov) between cited and citing authors.

![Semantic linkage network, a fragment](image)

**Figure 2. Semantic linkage network, a fragment**

The system can notify an author about new citations of their papers, including their context. Well-known RIS ResearchGate, Academia.edu and others already do it. The system can also provide the author with links to PDFs, which cite their papers, where the new citations are highlighted/annotated and work as interactive elements or as a communication instrument.

If the system identified a user as the author of the cited paper and this user clicks the in-text citation pointed to his cited paper, the system allows this user to express publicly or privately his reaction on how the citing author used/cited his research. According to the citation content, this reaction can be as simple as “agree/disagree”
or can provide explanations by the cited author about what was wrong with using his outputs, or how it could be used properly, etc. If the cited paper has several co-authors the system allows them to express their “agree/disagree” with a reaction of one of them.

The system itself also can initiate some direct communications between citing and cited authors using the sense of the citation content, such as citation polarity (author’s disposition toward a cited paper) or citation function (a purpose of citation) (Hernández-Alvarez and Gómez).

Citation Content Analysis

Citation content analysis, for which CyrCitEc already provides about 750,000 records of the right and left contexts for each in-text citation, has a lot of attention from researchers. Waltman, in his review of the traditional citation impact indicators, proposed different ways to improve the indicators, including taking into account “the context in which a publication is referenced (i.e., the sentences in a citing publication around the reference to a cited publication)” (Waltman 43).

In recent years, methods for analyzing the content of citations have been actively developed. Some studies (Zhang et al; Ding et al.) present concepts of content-based citation analysis (CCA), which addresses a citation’s value.

Practical experiments with the analysis of in-text citations (they are also called as the in-text references) on various sets of full text papers are also known. One of them identified verbs in citation contexts (Bertin and Atanassova, “Factorial Correspondence Analysis”) and later they characterized the different sections of articles in terms of the verbs that appear in citation contexts (Bertin and Atanassova, “Factorial Correspondence Analysis”). Another aspect of CCA is how references are distributed along the structure of articles, and the age of these cited references (Bertin et al.). Some authors analyzed in-text citations as functions of time, textual progression, and scientific field. They built characteristics of in-text citations in over five million full text articles (Boyack et al.).

Hernández-Alvarez and Gómez, in their survey of CCA, provided information about tasks, techniques, and resources, including such tasks as the citation polarity and function classifications.

The analysis of citation polarity/function has the potential to draw conclusions about the motives of authors in citing papers. Such analysis can also produce suggestions:
what exactly was used from the cited papers and why. In some cases, this information may be critically important to the authors of the cited papers and may help to initiate direct communication between them and the citing authors.

If CCA recognizes criticism and the system notifies the author of the criticized paper, it gives them an opportunity to correct mistakes and further develop their research.

If CCA informs the author about a positive impact of their paper, then the author can conclude how to develop their research to strengthen research results of other scholars who cited them.

In both cases, cited authors will benefit if they inform citing authors about their progress with the cited research.

As a result, citation networks will become true communication networks. When implemented, such a system should theoretically allow researchers to directly collaborate with each other without the mediation of the current publishing infrastructure.

However, it wouldn’t be limited to pairs of communicating researchers. It is an act of, at least, direct triple communications since, if we observe a citing author, they can enter in direct communication, not only with authors they cited but also with authors who cited them.

We should address the situation when researchers cooperate and coordinate their activity in a group that includes three parties:

• themselves (“producers”);
• authors whom they cite (“suppliers”);
• authors who cite them (“consumers”).

It means cooperation where “suppliers” and “consumers” also have direct communication and can directly affect each other’s activity. Since the group is “a collection of people committed to work jointly toward at least one group goal” (Randrup et al.), the goal of this group is obviously the creation of new scientific knowledge.

Towards a Global Scholarly Collaboration System

By providing direct scholarly communications for the participants of scholarly virtual cooperation, who have traditionally collaborated via publication exchange and cita-
tion, it becomes theoretically possible to create more favorable conditions for their actual collaborative creativity and the development of new scientific knowledge.

Let us consider a joint activity of scholars in small groups (i.e., laboratories, project teams, etc.), which are strongly based on direct communication between group members.

A group, and specifically a small group, is “a distinguishable set of two or more people who interact, dynamically, interdependently, and adaptively toward a common and valued goal/objective/mission, who have each been assigned specific roles or functions to perform” (Mathieu et al. 274). A key feature of small groups is a harmonization of activities on the principle of “all with everyone”.

Another important feature of small groups is the high variability of their environment and, as a consequence, the need for group members to quickly adapt to changing conditions. As Mathieu et al. underlined, “... in order to adapt effectively, team members must predict what their teammates are going to do and what they are going to need in order to do it” (Mathieu et al. 274).

Cooperation is generally defined as a “joint effort toward a group goal” and as a “concerted collaboration”, which is a collaboration with “no identifiable individual deliverables; only group deliverables, toward which members must contribute simultaneous efforts” (Randrup et al.).

Cooperation becomes “concerted collaboration”, if the participants can create a collective model of their activity and habitat. They need the collective model to “play” (simulate) and analyze the various possible options for cooperation. Such a collective model arises if the group members can share with the group their personal mental models and hence create a collective (team) mental model.

The basic idea of mental models is that humans, by their mental reflection, construct internal working models of the world:

When interacting with the environment, with others, and with the artifacts of technology, people develop internal mental models of themselves and the things with which they are interacting. These models provide predictive and explanatory power for understanding these interactions. (Badke-Schaub et al. 7)

The concept of a shared or collective mental model is defined as:
knowledge structures held by members of a team that enable them to form accurate explanations and expectations for the task, and, in turn, coordinate their actions and adapt their behavior to demands of the task and other team members. (Jonker et al.)

The basis of the collaboration mechanism, which works for a small group, is that of shared mental models. A background in this area includes research on the development of public institutions (Denzau and North), increase in the effectiveness of joint activities of people in a group (Mathieu et al.), interaction of people with software agents (Fan and Yen), environmental protection (Jones et al.), political activities (Richards), etc.

The sharing of group members’ mental models means that members inform each other about their intentions and possibilities regarding options of their joint activity. An aggregation of such information, received from all members, forms a choice area, which is available for analysis to each individual member.

Shared or team mental models are characterized as knowledge or belief structures that are shared by members of a team, which enable them to form accurate explanations and expectations about the task, and to coordinate their actions and adapt their behaviors to the demands of the task and other team members. (Badke-Schaub et al. 8)

In “Toward a Theory and Agent-Based Model of the Networked Economy”, we proposed a conceptualization of how a collaboration mechanism based on shared mental models works (Parinov).

An aggregation of members' shared mental models creates a collective mental model of the team. Team members interact with the collective mental model by taking information from it, playing with it (sorting out different ways of their cooperation), and by changing it. They can change in the model their personal information image. They can also propose new configurations of the group’s cooperation.

In figure 3 we illustrate these interactions by the example of forming a collective mental model (CMM) for a group of four members. Each member, by continuously exchanging information with others, forms and actualizes their own mental model of group cooperation and alienates it into collective model for decision making about the future group’s activity.
If group members have fixed in CMM a mutually acceptable configuration of their cooperation, then this configuration passes to the stage of practical implementation. The CMM, formed and set at this stage, is used by the members for the actual coordination of their practical activities.

There is a lot of literature on collaboration research, such as the Six Patterns of Collaboration, which suggests a conceptualization of collaboration as the following processes: generate, reduce, clarify, organize, evaluate and build commitment (Briggs et al., *A Six Layer Model of Collaboration*). The same authors also provide the Six-Layer Model of Collaboration, which includes collaboration goals, group work products, group activities, group procedures, collaboration tools and collaborative behaviors (Briggs et al.). Such research claims an intellectual foundation (Randrup et al.) for discussing computer-supported collaboration, collaboration support systems (Briggs et al., “Facilitator-in-a-Box”), integrated collaboration environment (Vindasius) and many others.
GLOBAL SCHOLARLY COLLABORATION

However, they do not address the basic collaboration instrument, that is, shared mental models. Without it, it is impossible to respond to the obvious research question: how can the mechanism of scholarly collaboration, which traditionally serves only members of small groups, work for the entire research community?

In this paper, we do not claim to give an exhaustive explanation of the question posed. Below we discuss what main tasks should be solved to make CMM a part of the social and technical research infrastructure. One of possible options is to implement CMM in RIS, like RePEc or Socionet. Based on services of these RIS we can implement CMM to develop tools for direct communication between citing and cited authors. Since these RIS for economics have about 70,000 identified authors linked with their papers, the proposed mechanism of collaboration will serve a significant part of this research community.

CMM within a RIS is already more than a “mental” model. In that case, it would be better to call it a collective information model (CIM).

On the subject of having CIM as part of a social and technical research infrastructure, we suggest the following main tasks:

• A synchronization of the individual mental models of the group’s members with their information images in CIM.
  This function can be realized with the help of an information system that collects and accumulates various data about the activities, intentions and capabilities of cooperative researchers. Tools need to be developed so as to allow a person to share their mental model with the information system, and also to ensure its continuous realization.
  The parameters of the implementation of this function are: the number of group members reflected in the CIM; the accuracy and completeness of the representations of their behavior, intentions and possibilities; and the speed and accuracy of updating changes in these data.

• Representation in CIM of the environment in which the group works, and a reflection of the changes taking place in this environment.
  An information system can also perform this function by collecting information about the environment where the group members cooperate. It is possible that in the near future this function will be implemented even more effectively in connection with the development of the “internet of things”.
  The parameters of the implementation of this function are: the size of the environment’s fragment reflected in the CIM; the accuracy and completeness of its representation; and the level of actualization of changes in the environment.
• Playing (making simulations) in the CIM with possible variants of group members’ cooperation. Various computer simulation models can be useful to implement this function, allowing a computer analysis of the best scenarios.

• Choosing the best variant for cooperation from many possible. With equal relations between group members (no subordination), the implementation of this function means that the members must negotiate. This requires a mechanism of collective decision-making.

• Realization of the chosen variant of cooperation in practice, including management functions over the joint activities of the group members. This requires a mechanism for collective management of joint activities.

Summing up the five points listed above, we can conclude that the implementation of the first three tasks are greatly influenced by a power of information technology, while the last two tasks also depend on social norms and rules, i.e., on institutions of cooperation.

**Conclusion**

If the small-group mechanism of collaboration is used on a larger scale, a cooperation in the research community may work in a more effective mode. Cooperative scholars can coordinate their activities faster by direct communication and that will foster their research performance.

Another consequence is that traditional publications and academic publishing infrastructure lose their monopoly as an instrument of global scholarly communication. This will create some challenges for sustainability of the global research community.

Another serious challenge is the rapid increase of communication among cooperative researchers resulting in a danger of information overloading for them. As Randrup et al. wrote: “Core insight with a significant negative impact on the performance of collaboration is which have been unveiled by research is cognitive overload and inertia. Individuals have limited attention resources”. It is a threat, but current research and development in areas like software agents, computer bots and artificial intelligence help humans cope with an increasing intensity of information flows and give optimistic perspectives for surviving in the coming digital era.
GLOBAL SCHOLARLY COLLABORATION

References


Bertin, Marc and Iana Atanassova. “A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard.” *BIR@ECIR*, vol. 1143, 2014, pp. 5-12.

Bertin, Marc and Iana Atanassova. “Factorial Correspondence Analysis Applied to Citation Contexts.” *BIR@ECIR*, vol. 1344, 2015, pp. 22-29.


Ding, Ying, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang and Chengxiang Zhai. “Content-Based Citation Analysis: The Next Generation of Citation Analysis.” *Journal of the Association for Information Science and Technology*, vol. 65, no. 9, 2014, pp. 1820-1833.


Neylon, Cameron. “Citation Metrics are Making Headlines, but What Does Citation Really Mean?” JISC blog, 8 December 2017, https://jisc.ac.uk/blog/citation-metrics-are-making-headlines-but-what-does-citation-really-mean-08-dec-2017.


Zhang, Guo, Ying Ding and Staša Milojević. “Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content.” *Journal of the American Society for Information Science and Technology*, vol. 64, no. 7, 2013, pp. 1490-1503.
The Value of Network Sustainability
Why we Join Research Infrastructures

Elisabeth Ernst
Max Weber Foundation – German Humanities Institutes Abroad
ernst@maxweberstiftung.de

Abstract: This paper develops the concept of network sustainability. To become and stay sustainable, distributed research infrastructures must satisfy present needs while at the same time be flexible and resilient to meet future requirements. With regard to this, it is not enough to merely build a resilient economic model and be technically viable. Research infrastructures that can understand, address and shape future needs to have a sustainable community network. Clear characteristics of a research infrastructure with a sustainable network are that partners gain access to other networks and interest groups, that knowledge, information and expertise is shared freely among partners, that the infrastructure increases partners’ visibilities and vice versa, and that partners are enabled to stay current and state-of-the-art. This is shown on OPERAS (open access in the european research area through scholarly communication), a research infrastructure for open scholarly communication in the social sciences and humanities, and its member the Max Weber Foundation, a German research institution.

Keywords: network sustainability, scholarly communication, open access, open science, social sciences and humanities.

Acknowledgements: The author wishes to thank Fabian Cremer, Michael Kaiser, and Judith Schulte (Max Weber Foundation – German Humanities Institutes Abroad) for their precious insights and suggestions. This paper is largely based upon the findings of the OPERAS Design Study, a deliverable of the OPERAS-D project funded by the European Commission in Horizon 2020.

Introduction

In the social sciences and humanities, scholarly communication is at the heart of researchers’ activities. Scholarly communication practices differ greatly from those
in other disciplines. An example of this is the format of the monograph, which reveals
other particularities such as the workflow, the way researchers collaborate, the cor-
relation between fieldwork and theory, and the way arguments are elaborated and
constructed (Giglia). The publication thus cannot be considered the last step of the
research process or merely as the output produced, but it is indissolubly linked to the
whole research process (OPERAS “OPERAS Design Study”).

The research landscape in the social sciences and humanities, however, is diverted
and fragmented into an environment of many small players who are highly specialized.
This is mainly due to multiple disciplines and sub-disciplines, as well as to different
languages (Mounier). This specificity in the social sciences and humanities should not
be considered a flaw but rather an adaption to the reality of the research landscape.
Nonetheless, the fragmented scientific community becomes disintegrated if the dif-
ferent players act in isolation and negatively impact one another (Mounier).

To tackle the situation, the European Commission has started setting up the Euro-
pean Open Science Cloud (EOSC), an infrastructure to support open science in
Europe (European Commission). At the same time, a research-driven approach to the
internationalization of scholarly communication in the social sciences and humanities
is needed. This approach can bring together researchers from multiple disciplines
and various infrastructures, using different tools and languages, and integrate their
research into the EOSC.

This paper explores how OPERAS (open access in the european research area through
scholarly communication), a research infrastructure that addresses these particular
challenges in the social sciences and humanities, implements network sustainability
and how such a sustainable network benefits and at the same time arises from its
partners. The Max Weber Foundation, an OPERAS core group member, is used as an
example. The paper provides an overview of the Max Weber Foundation and the OPE-
RAS research infrastructure and uses a network analytical approach to OPERAS as
a community network. It then investigates the concept of network sustainability and
presents concrete examples of how this is implemented within the research infra-
structure.

Max Weber Foundation

The Max Weber Foundation is a legal entity, funded by the German Ministry of Educa-
tion and Research, but nonetheless independent in its activities as a research organi-
zation. The foundation is comprised of ten humanities research institutes abroad. Its
goal is to promote research focusing on history, culture, economy and on the social sciences, and to foster a mutual understanding between Germany and the guest countries. The electronic publication platform perspectivia.net is the international, cross-epochal and interdisciplinary portal of the foundation (Max Weber Foundation).

The Max Weber Foundation closely works with national infrastructure partners, especially DARIAH-DE, to develop services at a national level. DARIAH-DE is a consortium developing and providing digital infrastructure and services to support research in the humanities and social sciences with digital methods and procedures (DARIAH-DE Consortium). With joint research projects and cooperation partners, such as the “Forum Transregionale Studien”, the Max Weber Foundation takes part in the internationalization of research in the social sciences and humanities. The Forum is a national German organization with partners from different research centers. It seeks to internationalize research by enabling scholars from abroad to work on transnational research topics as invited fellows or at conferences (Forum Transregionale Studien).

Without having any significant service provider for an information infrastructure in their own organization, the Max Weber Foundation relies on direct partnerships and national or international initiatives including libraries, computing centers, digital humanities centers and joint infrastructure projects. As a core group member of OPERAS, the Max Weber Foundation has a vital interest in shaping a sustainable research infrastructure and integrating research in the social sciences and humanities internationally.

**OPERAS – A Distributed Research Infrastructure**

OPERAS was born from a clear understanding that the specific challenges in scholarly communication in the social sciences and humanities have to be addressed in a common effort. The research infrastructure presently gathers 38 organizations from fourteen European countries as well as two international partners and is coordinated by a nine-member core group. It is coordinated from France by OpenEdition, an infrastructure dedicated to electronic resources in the humanities and open sciences, and Huma Num, the French infrastructure aiming at facilitating the digital turn in these disciplines. OPERAS members’ backgrounds are very diverse and cover the whole life cycle of a research publication: publishers and publication platforms, infrastructure providers and libraries, universities and research organizations. OPERAS has been supported by OPERAS-D (Design) and HIRMEOS (High Integration of Research Monographs in the European Open Science Infrastructure), two projects funded under Horizon 2020, the biggest research and innovation program of the
European Union. Those projects’ results are the backbone for the future services that the OPERAS research infrastructure will deploy and which are described below.

As key objective, the OPERAS-D project has prepared a design study that defines governance models, scientific and technical concepts for future services that the infrastructure will provide, and has established a roadmap to achieve these goals according to the requirements for long-term sustainability. The study’s main findings are:

- the need to consider scholarly communication as the heart of scientific research and not as one of its outputs, particularly in the social sciences and humanities;
- the fragmentation of the field;
- the need and the conditions for integration at European level (OPERAS Consortium, “OPERAS Design Study”).

The design study serves as a basis on which the future of OPERAS is built. The Max Weber Foundation’s formal role in OPERAS-D is to ensure a clear and efficient communication and dissemination of the project’s results.

The HIRMEOS project has aimed to integrate open access monographs into the open science ecosystem in a systematic and coordinated fashion. To improve interoperability between publishing platforms and referencing and indexing service providers, five sets of data and metadata are implemented on participating platforms: identification metadata, named entities data, certification metadata, annotation data and metrics metadata (HIRMEOS). The Max Weber Foundation, while not being one of the implementing platforms, formally participates in the communication and dissemination of HIRMEOS’ project results.

Using the two projects’ results, the research infrastructure OPERAS will deploy the following three services:

- The first service, a European discovery platform dedicated to social sciences and humanities resources, aims at overcoming obstacles to interdisciplinary practices by providing a platform which is dedicated to social sciences and humanities data, researchers and projects at European scale. This need, while identified in the OPERAS-D design study, was also acknowledged by the European Strategy Forum on Research Infrastructures (ESFRI), the European Commission’s most important instrument to develop the scientific integration of Europe, in its 2018 roadmap. ESFRI states that “[s]cholarly communication practices in the social sciences and humanities need to be reinforced in order to implement Open Science principles”
and has created a high strategic potential area for “digital services for open science” (European Strategy Forum on Research Infrastructures, Strategy Report). The OPERAS discovery solution will be integrated into the EOSC marketplace. OPERAS coordinates with EGI (an e-Infrastructure set up to provide advanced computing services for research and innovation), CESSDA (a consortium promoting international research cooperation and results), CLARIN (a European research infrastructure for language resources and technology) and DARIAH-EU (a network to enhance and support digitally enabled research and teaching across the social sciences and humanities) on setting up the general interoperability requirements and to add the OPERAS discovery solution to the EOSC catalogue, thus making it the social sciences and humanities data hub for the EOSC.

- The second service will be a **virtual ecosystem for research activation**, aiming at overcoming the obstacles that hinder the development of Citizen Science in the social sciences and humanities. More concretely, this service strives to support collaborative practices between researchers of the social sciences and humanities and other actors, such as journalists, artists, small-medium enterprises, civil society organizations, city administrations, etc., and to enhance the financial support to their projects.

- The third service will be the **redevelopment of the Directory of Open Access Books (DOAB) as a central service for OPERAS**. This includes the development of the system based on D-space as well as integrating the OAPEN Library, an OPERAS core group member and open access publication platform, into DOAB.

Not all OPERAS partners were actively involved in both of the finalized projects or in the planned projects. The Max Weber Foundation has been a project partner in both OPERAS-D and HIRMEOS, it will be involved in all new services, and is a core group member of OPERAS.
To ensure the continuity of OPERAS, the research infrastructure is currently applying to the ESFRI roadmap 2020/2021. Apart from this, it will further strengthen its strategic plan, prepare a business model as well as the transfer to a legal entity (to be developed until fall 2019), and produce a robust governance plan. Furthermore, OPERAS has created an active GO FAIR Implementation Network: the “CO-OPERAS IN”. GO FAIR, which is a bottom-up, stakeholder-driven and self-governed initiative, aims at implementing the FAIR (findable, accessible, interoperable, reusable) principles through Implementation Networks, so-called INs (GO FAIR). CO-OPERAS IN aims at

Figure 1. The OPERAS network
© Laetitia Martin, January 2018
bringing the FAIR principles into the research environment of the social sciences and humanities and building a bridge between these data and the EOSC (CO-OPERAS IN).

**OPERAS as a Community Network**

Social networks have increasingly been studied since the early 20th century. With the rise of social network analysis, an interdisciplinary approach that works on the premise that all social life is formed by relations and on the patterns that these create, scholars of all disciplines have started to systemically study social networks (Marin and Wellman). Within social network theory, a network is often defined as “a set of nodes (or network members) that are tied by one or more types of relations” (Wasserman and Faust). In principle, everyone and everything can be a network member; including organizations, companies, web pages, countries, or even staff or volunteer positions (Marin and Wellman).

In the case of OPERAS, the fact that partners who have signed an application form, have accepted the principles and procedures and have subsequently been accepted as members, can be seen as the nodes of the social network. Even though the term social network is commonly synonymously used for social media networks such as Twitter and Facebook, a social network is not necessarily an online network. To avoid confusion, however, I will in the following use the term community network for the network of OPERAS partners rather than social network.
An attempt at grouping OPERAS members is useful to understand the community network’s structure. Partners can be grouped according to the countries they are based in (see figure 2), according to their formal function in OPERAS (nine partners are core group members, see figure 1) or according to the type of their organization. A categorization according to the type of organization could involve:

- publishers and organizations that are dedicated to open access publishing (Association of European University Presses; Institute of Literary Research of the Polish Academy of Sciences; Lexis Compagnia Editoriale; Linguistics in Open Access;
While it is helpful to have a clearer understanding of what types of stakeholders are partnering in OPERAS, a grouping of the individual stakeholders within the network is not advisable. A categorization according to countries does not take into account the different functions a member has within OPERAS nor the various types of stakeholders that are members of the research infrastructure. The same holds true for categorizing partners according to their function in OPERAS, which does not include the national differences between stakeholders nor does it differentiate between the types of organizations. Lastly, a grouping according to members’ types of organizations creates the problem of multiple group memberships. The Max Weber Foundation, for instance, is comprised of ten humanities research institutes abroad. Yet, the foundation also operates a publication platform. It is thus not only a member of the group of universities and research organizations but at the same time also a quasi-publisher. Another example is Georg-August-University Göttingen, which is a public German university but is participating in OPERAS through their university library. Therefore, in a subsequent section, which introduces the implementation of network sustainability within OPERAS, I will not group the research infrastructure’s members but rather look at all partners as interdependent but individual nodes.

What, then, are the benefits to a network analytical approach to the sustainability of a research infrastructure? It is difficult to find literature that explicitly deals with the centrality of community networks for the sustainability of research infrastructures.
There is, however, research on the role of social networks for project stakeholder management. Provan and Kenis show that the form of governance has a huge impact on the effectiveness of a network and Chung and Crawford, for example, propose to use social network theory to identify stakeholders and improve project management. In the field of technology clusters and companies, there is research that shows that social network formation is crucial to the sustainability of technology clusters and economic activities in some regions (Casper) and an influential paper has shown that the success of Silicon Valley can largely be understood through network analysis (Saxenian). For research infrastructures, a network analytical approach takes the attention away from the “static” expertise and formally defined tasks of a single member, and draws it towards the “mobile” connections that this member has with other partners and to the work that is done “informally”. It thus recognizes the value of sharing expertise and knowledge and of the informal connections within research infrastructures (Marin and Wellman). I argue in the following that these are crucial aspects to the sustainability of a distributed research infrastructure such as OPERAS.

**Network Sustainability**

Concepts for sustainable research infrastructures in the social sciences and humanities have only recently entered the international agenda.

The Brundtland Report of 1987 triggered a worldwide discussion on sustainability and environmental concerns. The report defines sustainable development as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (United Nations, “Report of the World Commission”). It furthermore outlines three areas of sustainability: economic, social, and environmental sustainability. This widely used definition of sustainability, even if not clearly inclusive of research infrastructures, can be adapted to them. Research infrastructures also need to satisfy present requirements while staying flexible and resilient to meet future needs—needs which are not at all self-evident. Sustainable research infrastructures must therefore constantly evaluate how the requirements they serve are developing and adjust to them.

The 2030 Agenda has enhanced the Brundtland definition in 2015 with 17 Sustainable Development Goals (SDGs) and 169 targets—none of which include the social sciences and humanities or target research infrastructures (United Nations, “Resolution adopted by the General Assembly”). Goal no. 17, however, which asks for a revitalization of global partnerships for sustainable development, defines “multi-
stakeholder partnerships that mobilize and share knowledge, expertise, technology and financial resources” as an important systemic issue (United Nations, “Goal 17”).

Adapting the Brundtland definition and the SDGs to research infrastructures, three areas of sustainability arise:

- **economic sustainability** which guarantees the research infrastructure’s funding and the efficiency of its governance;
- **technical sustainability** that includes data storage and the resiliency of dynamic software;
- **network sustainability** which includes gaining access to other networks and interest groups, sharing knowledge and increasing each other’s visibility, and staying current and state-of-the-art.

When is a network sustainable? Granovetter, who in 1973 introduced the now famous idea of the strength of weak ties, argues that while it is obvious that strong ties are beneficial to a network—they share information readily and communicate frequently—weak ties enable the network to gain information and access to nodes that are more distant and therefore to information that is original (Granovetter). Applying this to a distributed research infrastructure means that the network needs a strong core as well as diverse and distributed partners. More recently, Krebs and Holley have identified five general patterns that can be observed in effective networks: nodes share common attributes and goals, they are diverse, there are several paths between nodes, the average path length is short, and some nodes are more important than others. They furthermore argue that because networks are often left unmanaged, nodes sharing similar attributes connect and close by, resulting in a lack of diversity and effective paths of communication within the network (Krebs and Holley). Effective community networks, they argue, arise with an active leader taking responsibility and forming a hub. The leader then usually connects to outside potential partners, making them part of the network and thereby extending it. In an efficient and sustainable network, the leader needs to encourage connections between the nodes according to the individual needs. Some interactions might lie dormant, but the paths need to be created (Krebs and Holley). This “hub-and-spoke” model, however, should only be a temporary structure as it places too much power and liability in the hub. The network becomes more resilient the fewer nodes include the hub for their connections. The transition of the leader from the weaver of the network to a facilitator of network weaving is crucial for a sustainable network (Krebs and Holley). While Krebs and Holley exemplify their research on economic networks, it is an important insight that is also useful when looking at distributed research infrastructures. Thus, a sustainable network in the context of a distributed research infrastructure does not
merely imply forming connections to exchange information during professional and
scientific conferences in the short run—although the importance of this should not be
underestimated—but to achieve long-term and large-scale integration and unification
as a basic service for researchers.

Network sustainability for research infrastructures is, although phrased diffe-
rently, included in the report on long-term sustainability of research infrastructures
published by the ESFRI Long-Term Sustainability Working Group. They argue that “a
robust long-term vision is the most important prerequisite in order to successfully and
sustainably build and operate” a research infrastructure (European Strategy Forum
on Research Infrastructures, ESFRI Scripta). This vision goes well beyond financing
mechanisms and business models for research infrastructures. The report suggests,
among others, to launch continued and practically-oriented initiatives to improve the
management of research infrastructures “through the exchange of best practices and
lessons learnt, and contribute to strategic planning, evaluation, and training” (Euro-
pean Strategy Forum on Research Infrastructures, ESFRI Scripta).

Taking into account the research that has been done on the sustainability of networks
and bearing in mind the specificities of a distributed research infrastructure, I will
show in the following part that a sustainable network enables a distributed research
infrastructure to understand future needs, to then address those needs, and finally
to also shape them. This is not only beneficial for the research infrastructure but also
for the partners individually, as shown on the example of the Max Weber Foundation.

Network Sustainability: Practical Implementation

The projects OPERAS-D and HIRMEOS have substantially increased the sustainability
of the research infrastructure OPERAS. OPERAS-D has developed a concept for eco-

donomic sustainability, which includes a governance model and a future roadmap for
the research infrastructure in the OPERAS Design Study. This will further evolve in the
course of the application for ESFRI 2020/2021. The HIRMEOS project has addressed
the technical sustainability of the services developed within OPERAS. While both,
economic and technical sustainability, are closely interlinked with network sustaina-
bility, the following seeks to isolate how network sustainability is achieved and why
this is not only beneficial for OPERAS but also for its partners.
Understanding Future Needs

A clear benefit for the Max Weber Foundation and other partners arises from gaining access to other networks and interest groups that they can rely on when organizing workshops or conferences or are in need of specific information or expertise. The community network connects the foundation with other interest groups, for instance national infrastructures but also European initiatives as e.g., CESSDA, CLARIN, DARIAH-EU, and OpenAIRE (a project promoting open scholarship). While some of these are formal partners of the OPERAS infrastructure (DARIAH-EU, for example, is a partner of the HIRMEOS project) the connection to other networks and interest groups takes place through other members. In most cases, more than one partner in OPERAS can connect other members to these networks and interest groups.

An example of this is the CO-OPERAS Implementation Network which helps in bringing the FAIR principles into the SSH research environment, particularly in the vicinities of OPERAS members.

At the same time, OPERAS benefits from the Max Weber Foundation’s knowledge of and connections within the German scientific community and the host countries of its institutes. This became evident in the course of the OPERAS-D final conference that took place in June 2018, which brought together different stakeholders from and outside of OPERAS and has fostered, among others, OPERAS’ relationship with PKP. Regardless of the conference’s topic, the selection of relevant speakers was only possible because of the OPERAS community network. In addition, the conference has granted access to a larger international community to all OPERAS partners. The same holds true for the HIRMEOS project, where partners have organized and conducted webinars and workshops together.

This reciprocal network sustainability ultimately ensures that OPERAS stays and becomes aware of future needs which can only then be addressed. Gaining access to other networks and interest groups is only possible with a network of diverse partners, the strong “weak ties.” It can also only be robust if there are several partners that can connect others to outside networks and interest groups.

Addressing Future Needs

The strongest benefits from research infrastructures with a sustainable network arise from sharing knowledge, information, and experience. OPERAS has set up working groups which have developed white papers on the topics of advocacy for open access, best practices, common standards, multilingualism, open access business
models, platforms and services, and research and development tools. The working groups represent groups of partners sharing the same goals but who have diverse backgrounds. While in the beginning, the OPERAS coordinator has acted as a hub for the research infrastructure, the working groups are a first step to facilitate network weaving for the core group members of OPERAS who act as contact points for the working groups (see figure 1). The working groups’ topics stem from subjects identified in the OPERAS-D and HIRMEOS projects and will increase further in importance once they become formalized as “special interest groups” in the legal entity that OPERAS pursues in setting up in fall 2019. The Max Weber Foundation is the contact point of the advocacy working group and thus unites the expertise of various OPERAS members in this area. The working groups enable the Max Weber Foundation to actively pass on its expertise while at the same time acquiring knowledge from other partners. The working group papers are a result of this process where sharing information with a network benefits all partners. This especially holds true for the identification of future projects that OPERAS partners can work on together, which is an essential part of the working groups.

Another example from the HIRMEOS project is the metrics collection tool. Javier Arias from Open Book Publishers describes this and the benefits for organizations with fewer resources in “Collecting Inclusive Usage Metrics for Open Access Publications: The HIRMEOS Project” (Arias).

Sharing knowledge, information and expertise aids OPERAS in becoming sustainable as it prepares the infrastructure in solving challenges that it has previously identified and addressing future needs. To sustainably share knowledge, information and expertise, however, communication needs to happen via short paths but be organized through partners that are especially committed to the research infrastructure.

Another aspect that supports OPERAS in addressing future needs is that the research infrastructure increases the visibility of the Max Weber Foundation as a research organization internationally and contributes to the reputation of its researchers. This is important as a main obstacle to publishing open access for researchers is the perceived lower reputation (Ernst et al.). Increasing one’s own visibility and reputation outside the national context can thus help to meet this challenge. At the same time, OPERAS can more effectively address future requirements if it is visible, which becomes possible with more widely known partners.
Shaping Future Needs

More than merely addressing future needs, a sustainable community network supports partners of a research infrastructure in shaping the future landscape. The community network of OPERAS has for example supported the Max Weber Foundation in participating in an active discourse on open access and legal issues in Europe. It thus supports partners in staying current and state-of-the-art. More concretely, OPERAS has signed the Scholarly Publishing and Academic Resources Coalition’s (SPARC) open letter protesting against the reform of the European Union’s copyright directive, has led the Max Weber Foundation to sign the Jussieu Call for open science and bibliodiversity, thus supporting a flourishing and diverse academic publishing landscape, and has produced a common Declaration on the Plan S Implementation Guidance.

Another example stems from the HIRMEOS project, which enables the platforms that directly benefit from its services to develop and provide their own services within OPERAS and to integrate into the European Open Science Cloud. The Max Weber Foundation is a project partner of HIRMEOS without being one of the implementing platforms, but considers the implementation of the services in the future. The network value for the foundation arises from actively participating in the shaping of future services and the processes of standardization. The latter is urgently needed for the foundation as it specializes in transnational and transregional research projects.

Staying current and state-of-the-art is a prerequisite for partners of a distributed research infrastructure to actively shape future needs. This only becomes possible if partners of a research infrastructure share some common attributes and goals but are diverse enough to influence the landscape from different angles.

Obstacles to Network Sustainability

Implementation of network sustainability does not always work smoothly. Some of the aspects described previously can be achieved through the management of the research infrastructure and the business model. This includes, for instance, creating a core group to facilitate the network weaving of strategic partners, implementing regular video conferences to assist a continuous communication, and forming working groups to connect partners with similar goals. Also the formal requirements of funded projects supporting a research infrastructure contribute to a sustainable network.
Yet, the intrinsic commitment of partners involved is crucial. It stems on the one hand from the benefits of network sustainability for themselves and on the other hand on the sustainability and thus effectiveness of the infrastructure it creates. This causes a virtuous circle where partners’ commitments stem from the direct benefits of network sustainability, which increases the network sustainability, which in turn intensifies the advantages mentioned before. However, while the benefits are huge, they often do not show immediately. This can lead to a lower level of dedication than desired. Reasons for this include time constraints, a lack of labor force, money constraints, and a misjudgment of the overall work that participation in a distributed research infrastructure requires.

While this lowers the overall effectiveness of a distributed research infrastructure, it does not necessarily impede a sustainable network that can understand, address and shape future needs. The level of commitment does not have to be equally high for all partners at all times because a sustainable network has several paths between the nodes. One can lie dormant and others can be used in its place. Yet, oftentimes research infrastructures do not have several paths between the nodes. OPERAS has, for instance, only one partner in some geographical regions (see figure 2) and within the core group there is only one partner each who is the responsible contact point for a working group (see figure 1). If the level of commitment of one of these partners is low, the research infrastructure’s network sustainability risks impairment.

Another problem is the length of the average path between nodes, which should be as short as possible in a sustainable network. In the case of OPERAS, OpenEdition, which has in the beginning acted as the hub of the research infrastructure, began to enable the core group members to act as network weavers themselves. Examples of this are the working groups and the setting up of the CO-OPERAS Implementation Network, which has been the initiative of especially one core group member (UniTo) who then succeeded in setting up the IN together with the other OPERAS members.

The OPERAS Coordination Team, a team comprising the OPERAS coordinators from OpenEdition and Huma Num as well as other staff from those two organizations and the communication officer from the MWS, has been accepted to pursue the Executive Master in Management of Research Infrastructures at the University of Milano-Bicocca. Yet, the before mentioned examples do not shorten the communication paths of all members of the network equally, especially if they are only involved in one of the working groups. A network with long communication channels quickly becomes inflexible and unable to perform its function of understanding, addressing and shaping future needs.
Not all these obstacles will necessarily impede network sustainability. Yet, every research infrastructure is well advised to take these possible risks into account.

**Conclusion**

In order to become and stay sustainable, distributed research infrastructures must satisfy present requirements while staying flexible and resilient to meet future needs. This means that it is not enough to only consider their economic model and technical viability. Research infrastructures with a sustainable network can understand, address and shape future requirements through granting their partners access to other networks and interest groups, through sharing knowledge, information and experience and increasing each other’s visibility, and by enabling their partners to stay current and state-of-the-art. They can only do so if partners share common attributes and goals but are diverse at the same time. Communication between members of a research network has to follow short paths while some partners (a core group) are more important to the communication within the research infrastructure than others. Finally, there needs to be more than one path between partners to make a research network sustainable.

While some of these aspects can be achieved through the management of the research infrastructure and the business model, most derive, to a large extent, from an intrinsic commitment of the partners involved. The level of commitment does not have to be high at all times if the research infrastructure has more than one path of communication between its partners. However, this is not yet the case with many partners in OPERAS, making the research infrastructure vulnerable. In addition, the length of the average path of communication between partners should be as short as possible in a sustainable network. OPERAS has shortened the lengths of the communication channels by establishing working groups but they are still long for members of only one working group. This risks inflexibility and therefore the sustainability of the network.

Network sustainability deserves permanent efforts. While funded research projects are clearing the path for a sustainable community network, stakeholders of a research infrastructure need to actively engage. Shared services, such as conferences, trainings, and advocacy for open science require constant exchange of information. Projects may serve as development sprints and content providers, yet keeping track of new developments and evaluating techniques is a permanent task for all partners and goes beyond every project’s lifetime. While the project OPERAS-D has identified current challenges for scholarly communication in the social sciences and humanities,
a sustainable research infrastructure needs to constantly reevaluate the status quo. The solutions, particularly for unified services such as certification, discovery, and research for society, require continuous adaption to future needs. To be able to also shape those future needs, a sustainable research infrastructure must actively participate in a discourse in the respective field. Solutions to the challenges distributed research infrastructures face can for that reason only be sustainable if addressed by an international consortium.

References


Abstract: The current discourse around Open Science has tended to focus on the creation of new technological platforms and tools to facilitate sharing and reuse of a wide range of research outputs. There is an assumption that once these new tools are in place, researchers—and at times, members of the general public—will be able to participate in the creation of scientific knowledge in more accessible and efficient ways. While many of these new tools have indeed assisted in the ease of collaboration through online spaces and mechanisms, the narrowness of how infrastructure is imagined by open science practitioners tends to put the use of technology ahead of the issues that people are actually trying to solve, while failing to acknowledge the systemic constraints that exist within and between some communities. Drawing on an analytical framework grounded in Black feminist intersectionality (Noble, “A Future”), this paper highlights the need for more inclusive knowledge infrastructures, particularly in the context of fair and sustainable development. Three case studies from the Open and Collaborative Science in Development
Network (OCSDNet), are outlined in order to illustrate the importance of moving beyond a definition of infrastructure as merely a technical or physical entity. These cases, arising from research conducted in South Africa, Brazil and the Caribbean, demonstrate how more sustainable and nuanced forms of collaboration and participation may be enabled through broader understandings of knowledge infrastructures. This paper further argues that leveraging the feminist concept of intersectionality when conceptualizing the development of knowledge infrastructures could be one way to move from narrow assumptions about standardized knowledge “users” towards more inclusive reimaginings of how a plurality of knowledge can be produced and shared via networked technologies.

Keywords: inclusive knowledge infrastructures, open science, collaboration, research tools, participatory design, intersectionality.

Introduction

When one considers the concept of infrastructure, a generic image of bridges, sewers, water pipes and electricity cables tends to come to mind. These large, physical channels, tools and structures lay the groundwork required for many humans to live safe and efficient lives. When functioning as designed, infrastructures tend to be rendered invisible for most users (Star). It is only when these physical entities break down that one tends to become cognisant of their contribution to daily life.

Between 2015 and 2018, the City of Cape Town experienced one of the worst droughts in recorded history, causing government officials to initiate massive water-usage restrictions to city residents and businesses. Citizens, media and experts began to calculate and speculate about when Day Zero would occur: the ominous future date in which all of the city’s usable water supplies would be depleted, should seasonal rainfall patterns fail to return. Loose plans were made for the establishment of emergency water distribution points, whereby all of the city’s four-million residents would be allocated a ration of 25 litres of water per day.

The threat of having no readily-available water for drinking, flushing toilets, bathing or cleaning was a shock to the wealthy (and predominantly white) Capetonians living in well-serviced and formal communities. A political name-and-blame media discussion ensued, whereby government officials were accused of incompetence and corruption for their presumed failure to upgrade the city’s deteriorating water infrastructure. Wealthier residents began stock-piling bottled water, drilling private boreholes and
purchasing water storage tanks in an effort to harvest any rainwater for home usage. Swimming pools lay stagnant and many would-be tourists altered their travel arrangements to avoid the inevitable fall-out of a city adjusting to no water. Amidst the chaos, the voices and lived experiences from Cape Town’s many informal communities, often ignored in public discourse, began to emerge through various social media channels: “Poor black people have lived with no running water for many years but we’ve never seen any day zero campaigns” (@Mthunzi_zn).

The Cape Town example highlights that while infrastructure might appear as a seemingly neutral groundwork for human activities, a deeper analysis tends to reveal highly political nuances, with significant insight regarding divisions of power and privilege within a given context. In this case, while rich and middle-class residents enjoy the invisibility of (usually) well-functioning water-delivery infrastructures, for poor communities, this gap in service delivery is an everyday feature in their lives. In short, infrastructure (in this case, publicly funded infrastructure), does not serve all users equally.

Moving beyond discussions of physical nuts-and-bolts infrastructure, a similar critical analysis can be applied to existing knowledge infrastructures. In the field of open science, advocates have sought to “transform how we think about the collection, dissemination and value of data, the collaborative potential in science, and the public character of research” (Bezuidenhout et al., “$100 Is Not Much”). Action and discourse have tended to focus on the creation of new technological platforms and tools to facilitate sharing and reuse of digital information. There is an assumption that once these virtual infrastructures are in place, researchers and other collaborators will be able to participate in the creation of scientific knowledge in more equitable and efficient ways.

However, as the Cape Town example illustrates, it is short-sighted to assume that just because the invisible technical groundwork is in place for virtual collaboration, that users of this knowledge infrastructure will benefit equally from its installation. Instead, from a development perspective, this paper argues that it is necessary to reflect critically on who is being both included and excluded in the design and use of knowledge infrastructures. For the purposes of this paper, we use the term “inclusive knowledge infrastructures,” defined as the tools, platforms, networks and other socio-technical mechanisms that deliberately allow for multiple forms of participation amongst a diverse set of actors, and which purposefully acknowledge and seek to redress power relations within a given context. While many virtual knowledge infrastructures have certainly assisted in facilitating collaborations through online spaces and mechanisms, there is less attention on the socio-political implications
that transpire in terms of how and by whom such tools are used, and whose world views are being packaged and sold as most legitimate through these platforms.

Findings from the Open and Collaborative Science in Development Network (OCSD-Net) demonstrate the importance of a critical reflection on inclusivity in the design and operationalisation of knowledge infrastructures in order to foster opportunities for fair participation in scientific knowledge creation. Building on Safiya Umoja Noble’s call for intersectional approaches to the study of the materiality of digital infrastructures (“A Future”), this paper presents three case studies generated by OCSDNet research teams between 2015 and 2017, which together highlight relevant aspects of what to consider in the design of inclusive knowledge infrastructures. The feminist concept of intersectionality, which foregrounds awareness on the concurrent existence of racism, sexism and other forms of inequality, is important for those attempting to conceptualize the (re)development of more inclusive knowledge infrastructures.

**Literature Review**

In the field of development research and practice, there is often a tendency for funding agencies to push for more openness through technological research interventions, often for the sake of technology itself (Bezuidenhout et al., “$100 Is Not Much” 45). This type of technocratic “supply-side” thinking resembles critiques of the Information and Communication Technologies for Development (ICT4D) field (Krauss), whereby advocates tend to assume that people will inherently benefit from new forms of technical infrastructure, rather than focusing on what diverse communities are “demanding” from a development perspective (Chirumamilla and Pal). Indeed, even at the global level of development institutions (such as the UN), ICTs continue to be treated as “apolitical tools for development” (Singh 7).

On the other hand, a growing number of scholars in the fields of anthropology, science and technology studies, informatics and development have begun to recognize the importance of critically reflecting on the role of knowledge infrastructures as often-invisible, but essential groundwork for human achievement and empowerment (Kenner; Edwards et al.; Star; Kelty). This section begins by outlining several examples of the ways in which different researchers have defined and designed knowledge infrastructures within their contexts and the role of these infrastructures in the project’s success (or failure) in the long run. The intention here is to demonstrate that “knowledge infrastructure” is a loose but important concept to consider in the design of research agendas that are more inclusive, equitable and meaningful to
diverse actors. In what follows, we outline some of the growing literature related to knowledge infrastructures and intersectionality, and focus on several critiques of universality, as raised by feminist scholars of science investigating intersectional biases in digital technologies and platforms.

**Knowledge Infrastructure Literature**

Denisa Kera suggests that the emerging global movement around open hardware (in the form of technological tools, apps and hardware) has inspired an emergence of “geek diplomacy” whereby citizen scientists of both the North and South are inspired to create and share technical solutions to real-world challenges that are important to them. She argues that because of a focus on sharing within the movement, strong networks have emerged amongst and between Southern and Northern actors, with a side-effect being the creation of a critical-political body that is able to defy technical norms through “hacking.” In this case, the diversity of open hardware tools are both the products of citizen-scientist hackers, as well as the essential knowledge infrastructure that allow for networking, sharing and re-purposing between geographically diverse communities.

In a similar vein, Julia Elyachar observed that the mostly informal “communicative channels” between poor women in Cairo allowed them to gain information regarding their communities and contribute—in unobvious ways—to the economy and status of their families. The author suggests that these channels are an important form of social infrastructure, as vital for the city’s economy as physical infrastructures like roads and bridges. Just as well-functioning physical infrastructure tends to be largely invisible to most users (Star), women’s labor and the communicative channels forged through time and effort tended to be overlooked within a capitalist economy that privileges more obvious income-generating activities typically undertaken by male members of a household (Elyachar). Similar findings are echoed by Simone, who noted the importance of “people as infrastructure” (407) in the context of resilience and survival for the poor in Johannesburg. In these cases, the concept of knowledge infrastructure can be applied to a construct that is neither physical nor technical but largely social. Neylon suggests that social infrastructures have many of the same qualities as other forms of infrastructure: they are mostly invisible, form the groundwork and networks for other key livelihood activities, and demand time and skill to create and maintain.

Beyond grounded examples of knowledge infrastructures from the field, other researchers have acknowledged the assumptions and restrictions that ultimately resulted in the creation of inaccessible and unusable knowledge infrastructures. For instance, in
an interview, Christopher Kelty relays a critique of an open education resource (OER) called Connextions:

Connextions taught me that even if you build a really awesome technological system that completely blows your mind . . . it will mean nothing in a context where people are not looking for a mind-blowingly different way of creating what is essentially a textbook. If academics are content to write and teach in the way they always have, and if students are not really agitated about doing it differently, then there really isn’t any way around that inertia. Connextions, like a lot of technological projects, was too proleptic and not diagnostic enough: it imagined a world in which all sorts of problems were solved: automated markup, easily transformed documents, remixability of content, a centralized repository of freely available teaching modules. But these weren’t (and perhaps still aren’t) the problems most teachers face.

Kelty’s critique of technological infrastructures failing to account for the genuine needs of their intended users could also be applied to many examples of ICT4D work, where a “solution” is proposed to a so-called “problem” that is in fact a problem perceived by technology designers rather than the technology users, highlighting the difference between demand-side versus supply-side development thinking.

Other scholars, recognizing the limitations of current knowledge-sharing infrastructures, have developed their own iterations, including the Platform for Experimental Collaborative Ethnography (PECE) (Fortun et al.). PECE focuses on enabling online collaborative analysis, comparison, data re-use, and preliminary visualization. According to the design team, PECE is not only an investment in technical research infrastructure to support collaborative, experimental ethnography; it is also an experimental system in its own right, one designed to simultaneously conduct and query empirical humanities and creative social sciences. One of the ways it does this, for example, is by sustaining a running list of what is called substantive logics. In this case, diverse collaborators are encouraged to add their own “logics” (extending from their own experience and research), thus pluralizing and substantiating how a given project makes sense. PECE is just one example of new types of sociotechnical knowledge infrastructures that are being built and studied by researchers themselves, with the aim of supporting knowledge pluralism. In a similar vein, recognising the shortfalls of many MOOCs, which tend to systematise a rather one-sided form of knowledge replication and learning, Jacque Wernimont from FemTechNet employed feminist

1. ICT4D: Information and Communication for Development
2. MOOCs = Massive Open Online Courses, which are generally free and available for anyone to join online.
theories to develop Distributed Open and Collaborative Courses that recognize the complexities of learning situations through the collective design of local platforms (Brown et al.).

Finally, it is important to be cognisant of the larger landscape of policy discourse and analysis within discussions of knowledge infrastructures. In order to see changes in the way that knowledge infrastructures are conceptualized and built, these changes need to be embedded in policy discourse early on. Albornoz et al. (“Framing Power”) suggest that there is currently an over-representation of business and industry actors at the policy level who push for the integration of privately owned infrastructures, tools and services into scientific systems. These interests, and the consequent tools that are created, tend to develop new forms of exclusion, which have a particular effect on knowledge producers at the grassroots, who do not have access to resources and training to use them. If more equitable and inclusive infrastructures are to be built, the current consolidations of power and profits by business and industry actors within scholarly infrastructures must be closely tracked (Chen and Posada).

**Feminism, Open Science and Intersectional Technologies**

Scholars working on questions of race, class, and gender have long argued that the margins can be sites of oppression and resistance, where subordinated groups and individuals can cultivate reflexive perspectives (Collins, “Learning from the Outsider”, *Black Feminist Thought*; Crenshaw; Fanon; hooks). Rooted in Black Feminist Theory and Critical Race Theory, the notion of “intersectionality” was first introduced by legal scholar Kimberlé Crenshaw in her seminal essay published in the late 1980s. The theory has since been taken up widely across disciplines and diverse cases (Carbado et al.) including analyzing how bilingual youth navigate multiple inequalities when they translate for their immigrant parents (Kwon), studying how the European Union handles multiple inequalities at the structural level (Verloo), and understanding the experiences of underexplored identities in U.S. higher education (Mitchell et al.).

In developing “intersectionality,” Crenshaw was responding to the tendency within identity politics to overlook or silence intra-group differences, a dynamic repeated throughout anti-racist and feminist movements, to the detriment of black women. Focused on both the structural and political aspects of intersectionality with regards to rape and domestic abuse, Crenshaw highlighted the importance of engaging with issues like violence against women of color through an intersectional lens.

Leveraging the concept of intersectionality, over the past three decades, feminist scholars have continued to raise important critiques of science, debunking notions
of objectivity and universality and emphasizing their social construction (Knorr-Cetina; Harding, *The Science Question, Objectivity and Diversity*; Haraway, “Situated Knowledges”, “A Game of Cat”). Feminist scholars of science have made critical interventions in revealing the construction of race/sex/gender/sexuality differences—and their historical variations—in modern Western sciences, especially the life sciences and medicine (e.g. Lock; Schiebinger; Fausto-Sterling; Martin; Terry). For example, Madeline Akrich pushed forward understandings of how innovators, designers and promoters of technical devices constructed representations of their users and how, by inscribing those representations in the technical and organizational choices that they make, technology creators framed the boundaries of possible relations between users and devices. Such work on the history and social construction of technology has been extended more recently to look at digital technologies and platforms, emphasizing in-built assumptions and ideologies in seemingly neutral digital technologies (e.g. Hicks; Noble, *Algorithms of Oppression*; Eubanks).

This latest generation of feminist science scholars has been working to complicate dominant, neo-liberal discourses about digital technologies as tools for social empowerment. By redirecting attention to the multiple strands of less-visible realities, these scholars have begun to highlight the ways in which science and technology are imbricated in circuits of state power and global inequality. We seek to extend this work to the growing scholarship and policy discussions surrounding Open Science which overlaps with but is not limited to movements advocating for Open Access and Open Data. While networked technologies hold great potential to enable more egalitarian processes of knowledge making, Watson-Verran and Turnbull’s findings from more than twenty years ago remain relevant today: “Western contemporary technosciences, rather than being taken as definitional of knowledge, rationality, or objectivity, should be treated as varieties of knowledge systems” (116).

In other words, if knowledge infrastructures are to be (re)designed for greater inclusivity, then a key first step is to build on the decades of work produced by feminist scholars of science and to recognize the asymmetries of power and contexts out of which Open Science is emerging. We are keen to direct attention, as Noble (“A Future”) and others have done, to the contradictions apparent in neo-liberal discourse of openness as a tool for social empowerment in order to encourage greater dialogue about how we may be able to counter and recapture the concept through grounded practices.

Building on this scholarship, this paper seeks to investigate how knowledge infrastructures (many of which are assumed to be neutral or apolitical) may in fact replicate and reinforce the gendered, raced and other sociopolitical imbalances that
exist within existing systems of knowledge production. In this paper, we offer three specific cases of knowledge infrastructures from the global South that broaden our understanding of the concept and explore the mutual co-constitution between social and material infrastructures. The subsequent analysis in this paper highlights that knowledge infrastructures encompass more than just technical tools and apolitical hardware. Indeed, knowledge infrastructures should be recognised as contributing towards the configuring of power relations and possibly holding opportunity for enabling diverse communities to become involved in knowledge-creation processes.

**Methodology**

The Open and Collaborative Science in Development Network (OCSDNet) was an international research network run from 2014 to 2017 that sought to address the fundamental question of whether and how open science has the potential to contribute to the achievement of development goals and opportunities. The network was composed of twelve international research teams throughout Latin America, Africa, the Middle East and Asia, and from highly diverse disciplinary backgrounds. Over the course of two years, each team explored the challenges and opportunities for an open and collaborative science, and the potential of open science to facilitate fair and sustainable development (Canhos et al.).

In this paper, the OCSD Network Coordination team draws on data generated by and collected about the twelve projects. As part of the meta-analysis for the network, research teams were asked to regularly reflect on their original research questions and approaches, and to assess whether and why their research was progressing as expected or not. As the Network Coordinators, we challenged them to reflect on the barriers or resistance they encountered and how/whether they sought to overcome any challenges. In addition, the projects also submitted detailed technical reports at the end of years one and two. Each project produced a variety of blog posts to share emerging learnings within the network and beyond. The meta-analysis of these materials, conducted by the OCSDNet coordination team, included two rounds of coding and analysis amongst team members in Peru, Colombia, Canada, South Africa and the United States. The first round of analysis entailed the coding of project-specific documents, while the second round included an analysis of insights gained across projects. See Chan et al. for further details on the methodology.

3. It is particularly important here to highlight the tension between setting up more decentralized infrastructures which are contextualized and aimed to meet user needs versus highly centralized system such as MOOCS which are controlled by already powerful actors. The latter appears increasingly to be subjected to control through the imposition of particular standards as forms of “soft governance” (Berg et al.; Chen and Posada).
Conceptual Framing: An Intersectional Approach to Knowledge Infrastructures

While feminist theory and practice have been widely influential in many disciplines over the past three decades, this work has had little noticeable uptake within the growing field of Open Science. In this paper, we leverage the sensitivities to power, relationality, intersectionality and context provided by this work in order to explore how we might begin to analyze and understand very different situations and practices with regards to open science in diverse contexts of the global South. The following section outlines three case studies drawn from the work of OCSDNet research teams. These cases were selected from amongst the twelve projects as they are best able to illustrate the types of diverse expressions of sociotechnical infrastructures that appear necessary to lay the groundwork for fruitful and just knowledge production. We draw out similarities and differences in how various actors chose to negotiate their respective scientific research processes under a paradigm of “openness.”

As noted by Edwards et al. (5), infrastructures are not “systems [with] end-to-end processes,” but rather “ecologies or complex adaptive systems,” in which a process of continuous learning and adaptation is (not always successfully) occurring. This imagining of infrastructure is particularly useful in the case of knowledge infrastructures, in the context of modern technology and networked collaboration. We suggest that the three cases illustrated in this paper demonstrate the importance of moving beyond a definition of infrastructure as merely a technical or physical entity and demonstrate how broader understandings of knowledge infrastructures may enable more sustainable and nuanced forms of collaboration and participation. The concept of intersectionality provides an important framework for assessing opportunities to imagine and realise knowledge infrastructures that are responsive to and inclusive of a diverse range of actors.

It is important to note that we offer these cases not as model templates of what we are calling “inclusive knowledge infrastructures” per se, but rather to highlight how the Open Science community needs to think more contextually about diverse local needs and actors in order to develop more inclusive infrastructures.

The following table summarizes each case’s contributions towards our understanding of the various points which policy makers, institutions and practitioners working on Open Science should take into account in the design and development of infrastructures for knowledge production and sharing.
Table 1. Summary of three OCSDNet knowledge infrastructure case studies

<table>
<thead>
<tr>
<th>Project Name</th>
<th>Inclusive Infrastructure</th>
<th>Key Actors</th>
<th>Why was an intersectional approach important?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Broker for Disaster Management in the Caribbean</td>
<td>The co-creation of a shared vocabulary in order to respond more effectively to natural disasters</td>
<td>Disaster response officials, university-based mobile technology developers, university academics</td>
<td>Diverse forms of governance, languages and institutional structures meant that collaboration around effective disaster-management response would be impossible unless a cohesive, shared vocabulary was designed and agreed upon by all actors.</td>
</tr>
<tr>
<td>Online Virtual Herbarium in Brazil</td>
<td>Collaborative development of research and data governance framework</td>
<td>Biologists, programmers, students, researchers, herbarium staff</td>
<td>While the project emerged with the goal of making herbarium data ‘virtual’ and ‘open’ to the public in Brazil, a SWOT analysis and intensive negotiations with institutions, technicians, etc. showed that ‘openness’ is not equally desirable or relevant for all actors.</td>
</tr>
<tr>
<td>Research Contracts with Indigenous Communities in South Africa</td>
<td>Co-design of a research contract that allows indigenous communities to define when, where and how their community knowledge is used by external researchers</td>
<td>Indigenous community leaders; law practitioners/researchers, feminist researchers</td>
<td>Indigenous communities in South Africa are diverse, with unique histories, languages and shared knowledge. Thus, there was a need for a flexible and inclusive research contract in order to secure the rights of communities during interaction with external researchers.</td>
</tr>
</tbody>
</table>

In what follows, we explain the context and details of the research projects summarised above, with a focus on detailing the various processes and changes that the projects undertook and why. Ultimately, we suggest that an intersectionality lens is important for planning, designing and assessing forms of inclusive knowledge infrastructures.
OCSDNet Case Studies: An Intersectional Approach to Knowledge Infrastructures

Case Study 1: Knowledge Broker for Disaster Management in the Caribbean

This project was led by a team within the Business School of the University of West Indies (MONA) in Jamaica. The team recognised that one of the core development challenges within the Caribbean Region was the ever-increasing scale of natural disasters, due to the combined effects of climate change and persistence of poverty, which leaves poor citizens most susceptible to injury and property damage. While most islands in the Caribbean are relatively small (in terms of population size and economic power), they experience the similar challenge of dealing with often-intense tropical storms. Given this shared challenge, the team recognised that there would be considerable advantage in collaboration between island states in order to respond more effectively to disaster situations. With the conception of OCSDNet in 2015, the team initially proposed the development of a mobile phone application, which they imagined would act as a virtual platform for collaboration between various disaster response units throughout the region.

However, as the team explored the potential for developing a mobile application for collaboration, they recognised that the more pertinent barrier to collaboration was the use of inconsistent terminology between actors and institutions of different languages, economies, political structures and geographical distances throughout the region. Thus, tapping into a more “demand-side” approach to development thinking, the researchers adjusted their work plan in order to further investigate what tools and/or processes could be designed in order to develop a cohesive disaster-response vocabulary across diverse disaster-management actors in the region. The result was the eventual creation of what the team calls a “knowledge broker” for disaster management in the form of a collaborative, virtual thesaurus that could be used by diverse technical actors.

Nonetheless, even with an adjusted research agenda, the team soon recognised that in order to develop more effective and inclusive knowledge infrastructure, a considerable amount of time and effort would need to be spent establishing networks with the correct actors including high-level politicians, technical staff and disaster management units. Hence, the network-building component of this project was perhaps even more important than the digital tool that the team eventually created.
In order to facilitate collaboration between the research team and regional disaster management groups, the project leads consciously presented their work in such a way so as to align it with the objectives and outcomes of the regional management bodies. In that way, the leads ensured that the partners could see the benefits of collaboration and sharing of resources. By translating the project into language already used by the partner organizations, the leads were able to insert “openness” into an already-existing set of goals and enable government stakeholders to realize how the rather abstract concept of “openness” could actually be applicable to a number of their existing working areas.

In short, the team recognised that a technological tool could not simply be created to fulfill the needs of a wide array of stakeholders representing disaster management sectors across various Caribbean countries. Moreover, instead of simply supplying a new and defined vocabulary for a diversity of actors, they directed their research actions towards network-building, alignment and consensus-making in order to collaboratively develop a disaster-management vocabulary that would be understood and easily accessible across countries and institutions of the region.

**Case Study 2: Online Virtual Herbarium in Brazil**

Given the rich biodiversity found in Brazil, large investments have been made into developing cyber infrastructures to support knowledge about Brazilian biodiversity (Barjak et al. 2013). In 2009, the Center for Reference in Environmental Information (CRIA), a non-profit Brazilian civil association, established the Virtual Herbarium of Flora and Fungi (INCT)—an e-infrastructure that provides the means for institutions to share their species data within a structured system responsible for data storage, search, retrieval, and visualization. The INCT documents, stores, disseminates, and increases the knowledge base on the diversity of plants and fungi of Brazil. Besides making 5-million data records and 900,000 images openly available, CRIA has developed new visualization tools to easily produce maps and graphs, and compare images, thereby enabling cyber taxonomy. In 2018 (as of April 4, 2018), there are 472 datasets on the database.\(^4\)

Prior to 2015, little attention had been given to analyzing the usage of INCT data as well as any incentives or barriers for openly sharing data through participation in the virtual herbarium network. But with OCSDNet support from 2015 to 2017, CRIA developed a project to critically assess network usage, evaluate whether the e-infrastructure was facilitating collaboration amongst scientists,

---

4. See: http://inct.splink.org.br/ for more information regarding the virtual herbarium repository.
and to understand whether open data was aiding users to create new forms of knowledge to help solve local development challenges.

The team developed a SWOT (strength, weaknesses, opportunities, threat) analysis of the virtual herbarium and circulated it through a large-scale survey sent out to all users and data providers within the network. On the side of data users, the team found that over 90% of all users were Brazilian, which demonstrates the importance of having openly accessible virtual data available in order to support the creation of local knowledge. In addition, the majority of the users were students (of varying levels in their scholarly careers), but also included academics, local scientists, members of NGO’s and the general public.

Interestingly, responses from the data providers within the virtual herbarium offered far more nuanced findings to contribute to the structuring of the network. In particular, the team found that the various institutions providing data to the network had often-conflicting conceptions of openness, which contributed to their willingness or hesitancy towards contributing data to the network. For instance, in an email survey, some institutions responded that they would not want to share species data that had not yet been published (due to fear of having the information stolen, and hence not being credited for their work); while others were afraid of revealing certain types of data—such as the locational information of high-value endangered species, due to the fear of having those specimens physically poached; potentially leading to species extinction. The worries raised by smaller data providers suggest that openness does not necessarily always benefit all actors. Policies and infrastructures developed to promote greater scientific openness stand to gain by recognizing and building for diverse, intersectional categories of users and creators.

As a result of their findings, CRIA developed a formal but flexible system for data providers to participate in the network. By providing technical options to open or “hide” different types of data, the data providers could choose which data they were willing to provide, rather than an all-or-nothing deal. For instance, a data provider could choose to set the name and image of a particular species as open, but the physical location of the same specimen hidden to the general public. In this way, the technical infrastructure still provides important reference data for general users, while also protecting the providers (and potentially the species) from data “poaching.”

In summary, while the Virtual Herbarium infrastructure was almost purely technical in its conceptualisation prior to OCSDNet in 2015, the team came to realize that some of the most challenging aspects of creating a sustainable and accessible open access platform are the negotiations and models of collaboration between different
WHOSE INFRASTRUCTURE?

institutions (with varying resources and priorities), as well as the disciplinary and vocabulary differences between botanists and ICT technicians. In the end, the team recognised that the creation of inclusive policies, a formal system of governance and sustainable long-term funding are all required to maintain the online database. In their final report, the team wrote:

For users to be able to rely on information systems, it is crucial for them to operate with uninterrupted, long-term funding... E-infrastructures require long-term maintenance and constant development, continuous and dynamic evaluation and planning, and efficient governance models to assure continuity of the network and its services. (Canhos et al.)

Hence, similar to the findings of the Caribbean team from Case Study 1 above, this example reveals that while knowledge infrastructures may appear largely technical, the sociopolitical elements are most often key to making or breaking its usage, and negotiating its inclusivity.

Case Study 3: Research Contracts with Indigenous Communities in South Africa

South Africa is home to a highly diverse group of people—including many indigenous groups that have lived throughout the southern tip of the continent for thousands of years. Historically, many of these communities lived in small, nomadic groups and have relied on intricate, generational knowledge of their environments in order to survive in the often-harsh climatic conditions that characterise much of Southern Africa. In 2014, a research team consisting of representatives from Natural Justice—a legal-research NGO in Cape Town—and academics from the United States applied to OCS-DNet for funding. The team outlined a proposal that suggested developing research relationships with Indigenous South African communities in order to understand (and potentially “open up”) local knowledge that could be important for understanding the impact and implications of climate change throughout the region. The team suggested that by “opening” Indigenous knowledge to the general public, fellow South Africans and others could benefit from generations of indigenous expertise in dealing with harsh climatic conditions.

However, as the team began to approach communities, they were met with immediate distrust. On previous occasions, some communities had engaged with researchers and had often been the victims of knowledge exploitation, whereby some researchers would extract local knowledge, for profit, without compensation for communities themselves. This is a phenomenon that has been increasingly documented
(Tuck and Yang; Smith) in other localities. Therefore, the very idea of harnessing and “opening” community knowledge on climate change for the benefit of outsiders was met with great resistance.

In response, taking a similar approach to the Caribbean team outlined in Case Study 1, the research team in South Africa decided to adapt the project to what indigenous communities were demanding from a development perspective, rather than what the researchers felt would be most appropriate. Recognising that communities were being exploited by external researchers, the OCSDNet team facilitated a process of developing a flexible and dynamic community-researcher contract, which would allow the community to negotiate, on (theoretically) more equal terms, with researchers and knowledge-profiteers with whom they might interact in future. As such, the idea of a community-researcher contract was determined to be relevant to implement for the project.

In this example, we view the community-researcher contract as an important example of an inclusive knowledge infrastructure. The research contract provides the framework, limitations and expectations for the sharing of knowledge from one party to another. Moreover, just as the Brazilian example has shown in Case Study 2, “openness” and the concept of “inclusivity” must be assessed critically: who is benefitting from open knowledge and who is exploited or excluded through these tools or structures?

The South African case study also reveals an example of the ways in which hierarchical knowledge structures might be challenged through the creation of new and inclusive knowledge infrastructures. For instance, in the design of the community-researcher contract, the team underwent intense negotiations with the American university partner, who sought to assert a universal application of research ethics and contracts for all research affiliates. As many academics are aware, these research-ethics contracts are often a one-size-fits-all model, structured solely by the university, and not by research participants themselves. The research team revised this top-down structure to create a more egalitarian contract that allowed the community to set the terms of exchange for their shared generational knowledge.

All in all, the research process and resultant infrastructure were deliberately strengthened through a series of back-and-forth discussions and negotiations over specific provisions of the community-researcher contracts. Through targeted efforts, the South African team refocused their research towards addressing community needs, rather than following the initial research agenda that would have contributed to the replication of existing power relations. Instead, by working with the community to
develop a community-researcher contract, the team helped to lay out the rights and responsibilities of all parties, to ensure that local knowledge is used fairly and with permission. Hence, this form of knowledge infrastructure has important implications for future development research and opportunities.

**Concluding Thoughts**

A majority of work and practice in the field of Open Science has tended to overlook the importance of social structures and systemic constraints in the design of new forms of knowledge infrastructures. In this paper, we highlighted the importance of ongoing negotiation and translation with a wide array of actors to develop infrastructures that are more inclusive to and mindful of the needs of diverse users and creators. In assessing the three OCSDNet case studies outlined above, we conclude that the concept of intersectionality provides a useful framework around which to understand the concept of “inclusive knowledge infrastructures.” Indeed, knowledge infrastructures must be mindful of the diversity of human needs, identities, abilities, experiences and forms of knowing. This finding also highlights the importance of cross-disciplinary learning. Open science practitioners can certainly benefit from the wide history of knowledge generated by decades of feminist scholars.

On the other hand, an intersectional framework also has its limitations. Within the cases outlined above, the meta analysis of the projects was limited by the forms and types of data that the research teams themselves generated. Some OCSDNet teams had an explicit background and interest in social science and development and hence were perhaps more attuned to the opportunity to assess the roles of gender, race and other complex identities within their projects. Other teams were from natural-science backgrounds and often expressed difficulty in understanding the benefits of such an approach, and the potential relevance to their respective projects. In the latter instances, we were left with scant material for our own analysis through formal reporting channels. On the other hand, one-on-one interviews with project leaders permitted an indirect opportunity to gain insight around the dynamics of each team, including how the research process was planned and implemented and what tactics were used and learned in order to develop knowledge infrastructures that were inclusive (or not) of diverse actors.

It is important to re-emphasise that the three case studies highlighted in this paper are by no means attempting to offer a template model for creating more inclusive knowledge infrastructures, nor are they even necessarily offering particularly ideal or successful examples of inclusive infrastructures. Instead, we have sought to highlight
how different histories, social and institutional systems, languages and cultural practices influence the way that different groups and individuals understand and desire “openness” in their respective circumstances. This, in turn, impacts the way that Open Science research processes are designed and ultimately the success or failure of more inclusive knowledge infrastructures. Intersectionality is offered as a way of critically assessing all of these themes and as a way towards realizing that both users and creators of infrastructures have complex, heterogeneous and (at times) conflicting needs and identities which should be accounted for in the design of new knowledge infrastructures.

Going forward, the OCSDNet coordination team seeks to understand how we might inspire and seed future research inquiries to more explicitly take notice of the ways that technology and research practices are intersectionally racialized and gendered, especially amongst research teams who may not have previous training or experience in feminist theory and practice. Indeed, this is an important question which we hope to see clarified in future iterations of the network.

Beyond OCSDNet, there is currently strong commitment from global institutions towards the development of physical infrastructure for sustainable human development. For example, the ninth UN Sustainable Development Goal (SDG) calls for the “building of resilient infrastructure, promotion of inclusive and sustainable industrialization and fostering of innovation” (United Nations). Infrastructure has also become an increasingly common buzzword in various overlapping academic fields. However, in spite of all of the attention on infrastructure, little attention has been brought towards the idea of developing more inclusive knowledge infrastructures, which we have defined in this paper as the tools, platforms, networks and other social and technical mechanisms that deliberately allow for multiple forms of participation amongst a diverse set of actors, and which purposefully acknowledge and seek to redress power relations within a given context. In an attempt to raise further attention to the importance of such infrastructures, this paper calls for the need to expand our thinking of what constitutes infrastructures, suggesting that a more direct focus on the creation of inclusive knowledge infrastructures is key for local development.

In this regard, Edwards et al. suggest that a more comprehensive focus on bringing “design thinking” into the creation of sustainable development infrastructures should be investigated, with a particular focus on encouraging user participation in the planning and design of tools and systems that are used on a regular basis. While this argument is important for the development of sustainable physical infrastructures as pursued by the SDGs, it is perhaps even more essential in the design of inclusive (and effective) knowledge infrastructures, such as those outlined within this paper.
Nonetheless, uncritical adoption of design thinking methodologies should also be avoided as critiques have already identified the possible replication of many similar issues raised against participatory research methodologies, such as the replication of status-quo power relations when not facilitated responsibly (Cooke and Kothari). In a forthcoming book chapter (Albornoz et al., “(Re)imagining ‘Openness’”), we suggest that a first step towards reimagining openness as a mechanism of resistance in pursuit of epistemic justice could be to assess who is absent in the design of open scholarly systems.

All in all, this paper and the case studies outlined seek to emphasise that there is no singular way of doing open science or producing new forms of knowledge. While discourse and practice in science tends to be dominated by a western-positivist world view that posits tools and infrastructures as seemingly neutral and homogeneous, this paper encourages a feminist approach to science and knowledge: one that recognises complex identities, histories, cultures and diverse forms of expression as invaluable for creating knowledge infrastructures that are deliberately and constructively inclusive.

References

@Mthunzi_zn. “.Poor black people have lived with no running water for many years but we’ve never seen any day zero campaigns.” Twitter, 6 February 2018, 7:40 A.M., https://twitter.com/Mthunzi_zn/status/960900897046958085


Availability of Cultural Heritage Structured Metadata in the World Wide Web

Nuno Freire
INESC-ID, IST, University of Lisbon, Portugal
(corresponding author)
nuno.freire@tecnicoulisboa.pt

Pável Calado
INESC-ID, IST, University of Lisbon, Portugal
pavel.calado@tecnico.ulisboa.pt

Bruno Martins
INESC-ID, IST, University of Lisbon, Portugal
bruno.g.martins@tecnico.ulisboa.pt

Abstract: In the World Wide Web, a very large number of resources is made available through digital libraries. The existence of many individual digital libraries, maintained by different organizations, brings challenges to the discoverability, sharing and reuse of the resources. A widely-used approach is metadata aggregation, where centralized efforts like Europeana facilitate the discoverability and use of the resources by collecting their associated metadata. The cultural heritage domain embraced the aggregation approach while, at the same time, the technological landscape kept evolving. Nowadays, cultural heritage institutions are increasingly applying technologies designed for the wider interoperability on the Web. This paper presents a study of the current application by cultural heritage data providers of technological solutions in use for making structured metadata available for re-use in the Internet. We investigated the use of both linked data and technologies related with indexing of resources by Internet search engines. We have conducted a harvesting experiment of the landing pages from websites of digital libraries that participate in Europeana and collected statistics about the usage of these technologies. These technologies allow for representing structured data within HTML, or for structured data to be referred to by links within HTML or through HTTP headers capabilities. We conclude with a discussion of future work for establishing a solution for cultural heritage aggregation based on the current situation and the available technologies.
Keywords: metadata, cultural heritage, search engines, linked data, World Wide Web.

Acknowledgements: We would like to acknowledge the supporting work by Antoine Isaac and Valentine Charles, from the Europeana Foundation, for their reviews and discussions regarding our work.

This work was partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and by the European Commission under contract number 30-CE-0885387/00-80.

Introduction

In the World Wide Web, a very large number of online cultural heritage digital resources is made available through digital libraries websites. These digital resources may consist of digitizations of physical objects (e.g., books, paintings, manuscripts) or born-digital objects (e.g., e-books, photographs). The discoverability of these resources through Internet search engines is still a challenge. Many cultural heritage resources are not of a textual nature (e.g., images, video or sound). Those that are textual often lack machine readable full-text, of which search engines are highly dependent, because they consist of digitized images where the application of optical character recognition (OCR) was not performed, due to lack of funding or availability of a mature OCR technology (e.g., for manuscripts or early printed materials). For discoverability, cultural heritage institutions have always relied on the creation of data records describing the resources.

These descriptive records are the basis for accessing and retrieving the resources through each institutional digital library website, which are specifically built for retrieval of this kind of data. The existence of many individual digital libraries, maintained by different organizations, brings challenges to the discoverability and usage of the resources by potential users, making the adequate indexing of cultural heritage metadata in Internet search engines even more relevant.

Across institutions, the discoverability problem is addressed by an organizational architecture based on a central organization (a role often fulfilled by a cultural heritage institution, but not always). These organizations approach discoverability of the resources by collecting their associated metadata descriptive records. The central organization has the possibility to further promote the usage of the resources by
means that cannot be efficiently undertaken by each digital library in isolation. They typically provide Web portals that contain cultural heritage focused search engines, also specifically built for this kind of data records ("Dublin Core Metadata Element Set").

In the domain of cultural heritage, the data aggregation technologies used are not the same as for Internet search engines. OAI-PMH has been the embraced aggregation solution, since it is highly specialized in fulfilling the requirements for the aggregation of metadata datasets (Fielding et al.). However, the technological landscape around our domain has changed. Nowadays, with the technological improvements accomplished by network communications, computational capacity, Internet search engines, and semantic data interoperability, the motivation for adopting OAI-PMH is not as clear as it used to be in cultural heritage (Freire et al., "A Survey of Web").

In the last years, the cultural heritage domain has been able to create sustainable aggregation initiatives, with self-sustaining business models. Examples are Europeana, DPLA, DigitalNZ, Trove and Digital Library of India, which are collecting and providing access to the public digitized cultural assets from Europe, the United States of America, New Zealand, Australia and India, respectively. However, the costs related to the implementation of the technical solution for aggregation are high for data providers. For these aggregation initiatives, reducing the effort required for data providers would bring more participants to their networks and lower the overall costs, therefore increasing the sustainability of the whole network (Verwayen). In this context, if cultural heritage aggregators were able to re-use the technological solutions in use for indexing by Internet search engines, data providers could benefit from several advantages. In particular, it would give data providers the following motivations:

- For those already implementing these technologies in their digital libraries, the process for sharing their data with cultural heritage aggregators would become extremely simple.
- For those that do not yet have these technologies in use, implementing the technical requirements for cultural heritage aggregation would be more rewarding, since discoverability through Internet search engines would come as a valuable extra benefit.

This paper presents a study of the current application by cultural heritage data providers of technological solutions in use for making structured data (or metadata, in the cultural heritage context) available for re-use in the Internet. We investigated the use of both linked data and technologies related with indexing of resources by Internet search engines. We have conducted a harvesting experiment of the landing pages
from websites of cultural heritage digital libraries that participate in Europeana and collected statistics about the usage of these particular technologies. These technologies allow for representing structured data within HTML or allow for structured data to be referred to by links within HTML or through HTTP headers. An analysis and discussion of the collected statistics is also presented.

We conclude with a discussion, based on the outcomes of this study, regarding future work for establishing a solution for cultural heritage aggregation based on the current cultural heritage scenario and the available technologies.

**Related work**

Although the use of linked data in cultural heritage has been the focus of much research, most of the published literature addresses mainly the aspect of the publication of linked data and do not fully address how the common aggregation approach of cultural heritage can be based on the existing published cultural heritage linked data (Simou et al.; Hyvönén; Jones and Seikel).

The most similar work to ours is that of the Dutch Digital Heritage Network (NDE) and the Research and Education Space project (RES)\(^1\) (Meijer and Valk). NDE is a Dutch national level program aiming to increase the social value of the collections maintained by the libraries, archives and museums in the Netherlands. NDE is still an ongoing project, and its initial proposals are based on specific APIs to enable data providers to centrally register the linked data URIs of their resources (“High Level Functional Design”). The current proposal of NDE, by being based in its own defined API, does not yet provide a solution purely based on linked data.

The Research and Education Space project has ended in 2017 but its results are still available. It has successfully aggregated a considerable number of linked data resources from cultural heritage sources. The resulting aggregated dataset can be accessed online,\(^2\) but an evaluation of its aggregation procedures and results was not published. From the technical documentation available, we can see that RES managed to give significant steps in the specification of key tasks to enable the aggregation of linked data (McRoberts). Some tasks however were not fully specified by the end of the project, and no further information has been published afterwards.

---

1. [https://bbcarchdev.github.io/res/](https://bbcarchdev.github.io/res/)
2. [https://bbcarchdev.github.io/res/collections](https://bbcarchdev.github.io/res/collections)
Generic technical solutions have been proposed by others to enable the aggregation of linked data (Rietveld et al.). However, a standards-based approach has not yet been put into practice within cultural heritage.

The work presented in this paper is done in the context of the research activities being carried out within the Europeana Network,\(^3\) for improving the network’s efficiency and sustainability (Verwayen). Linked data has been identified in our past work as one of the technical solutions with application potential (“Dublin Core”). The work described in this paper is part of a series of experiments addressing several Internet technologies for this purpose (Freire et al., “A Metadata Aggregation”; Freire et al., “Evaluation of Schema.org”).

### The Experimental Setup

In our harvesting experiment from the landing pages of resources from Europeana data providers, we have harvested samples from 31 different sources. In order to setup this test sample, we used the Europeana Search\(^4\) and Record\(^5\) APIs. The Search API was used first to discover the existing data providers of Europeana and their collections. Afterwards, on a second set of requests, Search API was used to discover a list of records from each collection. In subsequent requests, made on the Record API, we requested the complete metadata records of a sample of records per collection of each data provider. A diagram of the complete process of the experiment is shown in figure 1.

At most, 100 records per collection were obtained. From these records we collected the URLs of the landing pages on the data providers’ digital libraries. The metadata records were obtained in the Europeana Data Model (EDM) format (“Definition of the Europeana Data Model”). The URLs were obtained from the EDM isShownAt, property of the ORE\(^6\) Aggregation\(^7\) element. In total, the sample comprehended URLs from 31 data providers, 609 collections and 52,866 resources.

---

3. The Europeana Network is a community of 1,700 experts with the shared mission to expand and improve access to Europe’s digital cultural heritage, in the organization they work for and/or by contributing to shape Europeana’s services.
5. https://pro.europeana.eu/resources/apis/record
6. ORE - Object Reuse and Exchange: Data Model.
7. http://www.openarchives.org/ore/1.0/vocabulary#aggr
We issued two requests on each of the 52,866 landing pages: one request for the human readable version in HTML and a second request for the machine readable representation of the resource using HTTP content negotiation (“Sitemaps XML format”). We then processed the responses and collected statistics on the usage of three possible ways that these digital libraries could be encoding the metadata descriptions of the cultural heritage objects: HTML5 meta tags, RDFa/RDFa Lite and RDF data (in any of the commonly used serialization formats). For the analysis of the HTML5 meta tags, we have excluded the standard HTTP tags, since they are not meant to provide any descriptive data regarding the content of HTML pages.
An additional aspect is addressed in the experiment—the data model or namespaces of the structured data encoded in the landing pages. In particular, we are interested in gathering statistics regarding the use of two data models: Dublin Core Metadata Elements (Lagoze et al.) and Schema.org⁸ (used by Google and several other companies). These data models are the most likely to be nowadays in use by cultural heritage institutions to represent the metadata of their resources.

**Results**

The totals responses obtained from the requests issued to the sample of 52,866 URLs are shown in table 1. None of the responses for linked data resulted in valid RDF. The most frequent response was the HTML page, instead of RDF, therefore hinting that HTTP content negotiation was not even supported. In some cases, an error “Unsupported content-type” was received, and for some sporadic cases a JSON response was received, but it was not in a JSON-LD form, therefore, no RDF triples could be obtained from them. Our visual inspection of some of these cases detected that the JSON data was under a specific format, probably defined by a particular JSON API of a digital library system.

In the responses to the HTML requests, we detected a total of 25,276 HTML pages containing some form of structured data: 14,407 pages with HTML5 meta tags and 10,869 with Schema.org data, encoded in RDFa, RDFa Lite or JSON-LD (table 1).

<table>
<thead>
<tr>
<th>Linked data requests (content negotiation)</th>
<th>HTML requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDF</td>
<td>HTML with HTML5 meta tags</td>
</tr>
<tr>
<td></td>
<td>HTML with Schema. org (in RDFa, RDFa Lite or JSON-LD)</td>
</tr>
<tr>
<td>o (no valid RDF responses)</td>
<td>14,407 out of 52,866 27% (from 17 Europeana collections)</td>
</tr>
<tr>
<td></td>
<td>10,869 out of 52,866 20% (from 6 Europeana collections)</td>
</tr>
</tbody>
</table>

Table 2 summarizes the usage of the HTML5 meta tags. Whenever meta tags were present on the HTML pages, at least one of the standard HTML5 meta tags was

---

⁸. http://schema.org/docs/about.html
in use. In some cases, meta tags with properties using prefixes were also present. Although none of the HTML pages specified the namespace of the prefixes in use (it can be done by using RDFa), some of the prefixes are well-known, and typically they refer to the following namespaces:

- “dc”—Dublin Core Metadata Element Set;\(^9\)
- “dcterms”—DCMI Metadata Terms;\(^10\)
- “og”—The Open Graph Protocol.\(^11\)

The prefixes “eprints” and “egms” prefixes where found as well, but we cannot be certain to which namespaces they refer.

Table 2. The rdf:type of Schema.org RDF resources present in the HTML pages

<table>
<thead>
<tr>
<th>Meta tag property prefix</th>
<th>HTML5 meta tags</th>
<th>Number of HTML pages</th>
<th>Number of distinct Europeana providers</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML5 standard tags</td>
<td>14,407</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Dc</td>
<td>3,783</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Dcterms</td>
<td>1,377</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Og</td>
<td>791</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>egms</td>
<td>701</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>eprints</td>
<td>100</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Regarding the use of Schema.org metadata, in table 3 it is shown the URIs of classes in use, which gives an impression of what is being described there. Schema.org can be used to describe many aspects related to the HTML page and its content. Therefore, although Schema.org data may be present in the HTML, it may not be describing the cultural heritage object. And, in fact, we observed the usage of instances of ListItem, BreadcrumbList, SearchAction, Website and ViewAction, in several cases, which indicates that the cultural heritage object metadata was not part of the existing Schema.org data.

---

11. http://ogp.me/
### Table 3. The rdf:type of Schema.org RDF resources present in the HTML pages

<table>
<thead>
<tr>
<th>Schema.org Classes used</th>
<th>Number of instances</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://schema.org/Thing">http://schema.org/Thing</a></td>
<td>9,770</td>
</tr>
<tr>
<td><a href="http://schema.org/ListItem">http://schema.org/ListItem</a></td>
<td>9,888</td>
</tr>
<tr>
<td><a href="http://schema.org/Person">http://schema.org/Person</a></td>
<td>555</td>
</tr>
<tr>
<td><a href="http://schema.org/BreadcrumbList">http://schema.org/BreadcrumbList</a></td>
<td>509</td>
</tr>
<tr>
<td><a href="http://schema.org/SearchAction">http://schema.org/SearchAction</a></td>
<td>282</td>
</tr>
<tr>
<td><a href="http://schema.org/WebSite">http://schema.org/WebSite</a></td>
<td>282</td>
</tr>
<tr>
<td><a href="http://schema.org/Organization">http://schema.org/Organization</a></td>
<td>282</td>
</tr>
<tr>
<td><a href="http://schema.org/VideoObject">http://schema.org/VideoObject</a></td>
<td>229</td>
</tr>
<tr>
<td><a href="http://schema.org/ImageObject">http://schema.org/ImageObject</a></td>
<td>227</td>
</tr>
<tr>
<td><a href="http://schema.org/Book">http://schema.org/Book</a></td>
<td>199</td>
</tr>
<tr>
<td><a href="http://schema.org/ViewAction">http://schema.org/ViewAction</a></td>
<td>197</td>
</tr>
<tr>
<td><a href="http://schema.org/Article">http://schema.org/Article</a></td>
<td>127</td>
</tr>
<tr>
<td><a href="http://schema.org/VisualArtwork">http://schema.org/VisualArtwork</a></td>
<td>100</td>
</tr>
</tbody>
</table>

### Conclusion and Future Work

The results of this experiment make it evident that in spite of the numerous activities in cultural heritage for making available linked data, reaching it through automated means based on the WWW (i.e. the Web of Documents) is no yet feasible. The digital object entry pages, whose links are sent to Europeana, could not be automatically linked to their respective linked data representations, since they did not support linked data through content negotiation, and the structured metadata we found encoded within the HTML pages was very limited, or even non-existent in the majority of cases.

In order to make use of cultural heritage linked data for metadata aggregations, less automated approaches need to be employed to discover, link and adapt the aggregation systems to each dataset of the participating CHI data sources (SPARQL end points, data dumps, etc.). Alternatively, aggregators may start to define the technical mechanisms for making linked data automatically discoverable, accessible and usable for aggregation.
Another aspect we also conclude from the experiment, is that it supports the beneficial value of cultural heritage aggregation initiatives, such as Europeana and DPLA, for promoting the discoverability of the cultural heritage objects through both the WWW and linked data. The activities of aggregators in the publication of open linked data are likely to be the most interoperable source of cultural heritage linked data currently available (Charles et al.). The results of the experiment provide further motivation for the development of Europeana’s activities towards Schema.org publication of its dataset and cultural heritage metadata in general (Wallis et al.).

Cultural heritage institutions are typically very dedicated to their digital libraries and the quality of the metadata that describes their holdings. From this study we conclude that cultural heritage institutions have been neglecting the availability of their structured metadata according to Web standards when defining the functional requirements of their digital libraries systems. Or, in those cases that this requirement is being considered, cultural heritage institutions need to make sure that they have a testing process to ensure that the structured metadata is properly published in machine readable form, since this kind of functionality cannot be tested by humans.

The next steps of our work will be to survey technologies of the Semantic Web, linked data and vocabularies for the description of datasets. We will analyze these technologies in search for a solution that will enable the aggregation of linked data in fully automatized ways or requiring very little human intervention. Table 4 shows a list of those technologies that we have identified at this stage of our work.

**Table 4. Technologies of the Semantic Web, linked data, and vocabularies for the description of datasets**

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linked Data Platform (Speicher et al.)</td>
<td>“Linked Data Platform (LDP) defines a set of rules for HTTP operations on web resources, some based on RDF, to provide an architecture for read-write Linked Data on the web.” (Speicher et al.)</td>
</tr>
<tr>
<td>VoID - Vocabulary of Interlinked Datasets (Alexander et al.)</td>
<td>“VoID is an RDF Schema vocabulary for expressing metadata about RDF datasets. It is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloging and archiving of datasets.” (Alexander et al.)</td>
</tr>
</tbody>
</table>
DCAT – Data Catalogue Vocabulary (Maali and Reikson)

“DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. Publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites.” (Maali and Reikson)

Schema.org

The Schema.org vocabulary defines classes representing Datasets and their distribution.

EDM Datasets Profile (“Europeana Dataset Profile”)

This profile defines the elements used to represent datasets ingested by Europeana. The profile is mainly intended to be used to disseminate dataset level information via the Europeana API.


These technologies may enable the aggregation of linked data in fully automatized ways or with little human intervention.

References


An Expertise Recommender System based on Data from an Institutional Repository (DiVA)

Milena Angelova  
Technical University of Sofia  
mangelova@tu-plovdiv.bg

Vishnu Manasa Devagiri  
Blekinge Institute of Technology  
vishnu.manasa.devagiri@bth.se

Veselka Boeva  
Blekinge Institute of Technology  
veselka.boeva@bth.se

Peter Linde  
Blekinge Institute of Technology  
peter.linde@bth.se

Niklas Lavesson  
Blekinge Institute of Technology  
niklas.lavesson@bth.se

Abstract: Finding experts in academics is an important practical problem, e.g., recruiting reviewers for reviewing conference, journal or project submissions, partner matching for research proposals, finding relevant MSc or PhD. supervisors, etc. In this work, we discuss an expertise recommender system that is built on data extracted from the Blekinge Institute of Technology (BTH) instance of the institutional repository system DiVA. The developed prototype system is evaluated and validated on information extracted from the BTH DiVA installation, concerning thesis supervision of researchers affiliated with BTH. The extracted DiVA classification terms are used to build an ontology that conceptualizes the thesis domain supported by the university. The supervisor profiles of the tutors affiliated with the BTH are constructed based on the extracted DiVA data. These profiles can further be used to identify and recommend relevant subject thesis supervisors.

Keywords: data mining, DiVA, expertise retrieval, knowledge management, natural language processing.
Acknowledgements: This work is part of the research project “Scalable resource-efficient systems for big data analytics” funded by the Knowledge Foundation (grant: 20140032) in Sweden.

Introduction

Finding experts in academics is an important practical problem, e.g., recruiting reviewers for reviewing conference, journal or project submissions, partner matching for research proposals, finding relevant MSc or PhD supervisors, etc. In this work, we discuss an expertise recommender system that is built on data extracted from the Blekinge Institute of Technology (BTH) instance of the institutional repository system DiVA (Digital Scientific Archive). DiVA is a publication and archiving platform for research publications and student essays used by 46 publicly funded universities and authorities in Sweden and the rest of the Nordic countries.\(^1\) The DiVA classification system is based on the Swedish Higher Education Authority (UKÄ) and the Statistic Sweden’s (SCB) three levels classification system. Using the classification terms associated with student M. Sc. and B. Sc. theses published in the DiVA platform, we have developed a prototype system which can be used to identify and recommend subject thesis supervisors in academy.

Related Work

The discovery of expertise is crucial in supporting a number of tasks. For example, finding an appropriate expert when one needs guidance on a subject matter, or needs to fill a vacancy based on relevant expertise, or needs to find research collaborators working in similar areas, etc. The methods used for developing systems that facilitate mining of expertise can be classified into two main categories: methods based on mining unstructured information and methods based on social networking sites.

Unstructured information includes emails, Web pages, wiki, reports, etc. Text mining tools are used to index technical terms from unstructured documents, which can be queried to identify subject experts. Some of the important tools using this information include email expertise extraction (e3) system (Kruwich and Burkey), Contact-Finder (Campbell et al.), MIT Expert Finder (Vivacqua), etc. A major limitation of these methods is the authentication of information.

\(^1\) http://www.diva-portal.org/
Today there are many social networking sites, some specifically for the scientific community, such as Research Gate and Nature Network. Experts have to feed in information about their subject expertise, domains, publications, credentials, etc. A major limitation of these methods is in the adding and updating of information.

A common drawback to both approaches is that they are based on non-peer-reviewed information provided by the user, i.e., they can be biased.

In the recent years research on identifying experts from online data sources, such as the DBLP library, Microsoft Academic Search, Google Scholar Citation, LinkedIn, PubMed, etc., has been gradually gaining interest (Abramowicz et al.; Balog and de Rijke; Bozzon et al.; Boeva et al., “Semantic-Aware”; Boeva et al., “Identifying a Group”; Hristoskova et al.; Jung et al.; Singh et al.; Stankovic et al.; Tsiporkova and Tourwé; Zhang et al.). An example is a Web-based biomedical expert finding system, proposed by Singh et al., which can be applied to identify subject experts and subjects associated with an expert. The system builds and maintains a big repository of biomedical experts by extracting the information about experts’ peer-reviewed articles that are published and indexed in PubMed. Two other interesting approaches using semantic matching and developing trust algorithms are those of Charlin and Zemel, as well as Osman et al. Charlin and Zemel have developed an automated reviewer assignment system that can be used to find suitable reviewers for conference papers. The assessment of reviewers’ expertise is based on their previously published papers and self-assessed expertise about the submissions. Osman et al. propose a trust model that calculates the expectation about an agent’s future performance in a given context by assessing both the agent’s willingness and capability through the semantic comparison of the current context in question with the agent’s performance in past similar experiences. A comprehensive survey on the state-of-the-art methods in expert finding and summary of these methods into different categories based on their underlying algorithms and models is presented by Lin et al.

Institutional repositories have been around for at least 20 years, but it is only since recently we find some discussions on how to create new services using text analysis or data mining in institutional repositories or library catalogues (Kerdprasop et al.; Lawrence; Okamoto; Tonon and Fusco) and even more rare are any evidence on actual resources based on these ideas.

The use of university repositories is not so common, in spite of their wealth of bibliographic metadata of both local scientific records and student theses records. There are papers describing mining the contents of scientific papers for more effective information discovery and selection (Saggion and Ronzano; Tonon and Fusco; Okamoto).
Other areas that are reported in the literature is personalized services for library user using data mining technology or book recommendation systems for library user based on user profiling (Ding; Jomsri; Cui et al.). These are the main uses of data mining in university institutional repositories that we can find being reported in the literature so far. One of the reasons for not exploiting institutional repositories on a greater scale could well be that they are often created, administered and run by librarians who generally are not research oriented and especially not towards computer science. Another reason might be that the data in the repositories often is entered by researchers and students themselves and therefore, in many cases, are not quality controlled. An elaborated discussion of this issue can be found in the section “Data”.

**Identifying Experts from Institutional Data Sources**

Many scientists who work on the expertise retrieval problem distinguish two information retrieval tasks: expert finding and expert profiling, where expert finding is the task of finding experts given a topic describing the required expertise (Craswell et al.), and expert profiling is the task of returning a list of topics that a person is knowledgeable about (Balog et al.). In the considered context we need to deal with both expertise retrieval tasks.

**Profiling of Expertise**

An expert profile may be quite complex and can, for example, be associated with information that includes: email address, affiliation, a list of publications, co-authors, but it may also include or be associated with: educational and (or) employment history. This information can be separated into two parts: expert’s personal data and information that describes the expert area of competence. Personal data is used to resolve the problem with ambiguity. This problem refers to the fact that multiple profiles may represent one and the same person and therefore must be merged into a single generalized expert profile. Namely, the process of merging expert profiles is driven by the calculation of the similarity scores between different entities composing the profile, e.g., expert name, affiliation, email address (Boeva et al., “Measuring Expertise”).

The data needed for constructing the expert profiles could be extracted from various Web sources, e.g., LinkedIn, the DBLP library, Microsoft Academic Search, Google Scholar Citation, PubMed. In our work, we have used data extracted from the DiVA repository to construct the tutor profiles and ontology model. Initially, the extracted data was divided into personal data and a list of subject terms. Personal data has entities, e.g., first and last name of each thesis supervisor, affiliation containing
information about university, faculty and department, and also the academic role of the supervisor. Each BSc or MSc thesis presented in DiVA is associated with a list of keywords that are selected and entered by the author in order to describe their thesis subject. We have used these keywords and the built ontology to describe the supervisors’ area of competence. The built ontology is a conceptual model of the domain of interest and it is used to attain accurate and topic relevant expert profiles. When a conceptual model is missing then, e.g., the Stanford part-of-speech tagger (Toutanova and Manning) can be used to annotate the different words in the text collected for each expert (supervisor) with their specific part of speech. The tagger also defines whether a noun is a plural, whether a verb is conjugated. The annotated text can further be reduced to a set of keywords (tags) by removing all the words tagged as articles, prepositions, verbs, and adverbs.

In view of the above, an expert profile is defined as a list of keywords (domain-specific topics), extracted from the available information, describing their expert area. The problem with ambiguity has been resolved using the extracted personal data. The expert profiles with the same names are candidates to be merged into a single expert profile. For instance, if the supervisors’ names and affiliations are identical then their expert profiles are merged into a single one. Most supervisors who are affiliated with BTH have not entered their email addresses in the DiVA repository. Therefore, we have not used the email address as an entity to resolve the problem with ambiguity.

**Expertise Similarity**

An important issue in an expertise retrieval context is to establish a way to quantify how well the area of expertise of an individual expert conforms to a certain subject or to estimate the expertise similarity between two experts. The calculation of expertise similarity is a complicated task, since the supervisor expertise profiles usually consist of domain-specific keywords that describe their area of competence without any information for the best correspondence between the different keywords of two compared profiles. One possibility to measure the expertise similarity between two expert profiles (or between a supervisor profile and a subject profile) is by taking into account the semantic similarities between any pair of keywords that are contained in the two profiles. Several semantic distance and similarity algorithms can be applied within structured terminological resources (Leacock and Chodorow; Rada et al.; Wu and Palmer). In this case, the algorithms count the number of edges (links) between the two keywords (classification terms) in order to compute the relatedness of these terms.
In the current study, we use a modified Wu and Palmer measure, proposed in Manjula Shenoy et al., to calculate similarity between two ontological elements. In the considered context all expert profiles are described by a list of domain specific terms, where they are an expert. Assume that each expert profile \( i \) is described by a list of \( p_i \) keywords. In our experiments, we use the modified Wu and Palmer measure to calculate the semantic similarity between two keywords. Manjula Shenoy et al. propose an improved version of the standard Wu and Palmer measure. The modified Wu and Palmer semantic similarity measure \( s \) is defined as follows:

\[
s = \frac{(2N \cdot e^{-\lambda L/D})}{N1 + N2}
\]

where \( L \) is the shortest distance between two given concepts, \( D \) is the depth of the ontology, \( N \) is the distance from the least common ancestor to the root, \( N1 \) and \( N2 \) are respectively, the distances from the two considered concepts to the root, and \( \lambda \) is a coefficient that is 0 when the concepts are in the same hierarchy and 1 when they are neighborhood concepts. Namely, the authors consider how far the two considered concepts are semantically and where they are located in the ontology. The modified Wu and Palmer measure (noted by \( s \) herein), finds the shortest path between two concepts in the ontology tree and the depth of the whole ontology. Then the expertise similarity \( S_{ij} \) between two expert profiles \( i \) and \( j \) \((i \neq j)\), can be defined by the arithmetic mean of semantic similarities between the corresponding keywords, i.e.

\[
S_{ij} = \frac{1}{p_i \cdot p_j} \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} s(k_{il}, k_{jm})
\]

Where \( s(k_{il}, k_{jm}) \) is the semantic similarity between keywords \( k_{il} \) and \( k_{jm} \) calculated accordingly to the definition given above (Manjula Shenoy et al.).

**Expert Finding**

The experts finding task can be viewed as a list completion task, i.e. the user is supposed to provide a small number of example experts who have been used to work on similar problems in the past, and the system has to return experts with similar areas of competence. In our context, the user can be a university student, a director of a study program or other administrative staff and the returned experts are recommended subject thesis supervisors.

Another possibility is to present the domain of interest by several preliminary specified subject categories and then the available experts can be grouped with respect to these categories into a number of disjoint expert areas (clusters) by using some clustering algorithm, e.g. ones developed in “Semantic-Aware” and “Identifying a Group”
AN EXPERTISE RECOMMENDER SYSTEM BASED ON DATA FROM AN INSTITUTIONAL REPOSITORY

(Boeva et al.). In the considered context each cluster of experts can itself be thought as the expertise area of any expert (supervisor) assigned to the cluster. In this case, in order to select the right individuals (thesis supervisors) for a specified task the user may restrict their considerations only to those experts who are within the cluster that is identical with (or at least most similar to) the task’s subject. The specified subject and the expert area can themselves be described by lists of keywords, i.e. they can be compared by way of similarity measurement.

Description of the Prototype System and Results

The developed prototype system has been evaluated and validated on information extracted from the BTH DiVA installation, concerning thesis supervision of researchers affiliated with BTH. The extracted DiVA classification terms are used to build an ontology that conceptualizes the thesis domain supported by the university. The supervisor profiles of the tutors affiliated with the BTH are constructed based on the extracted DiVA data. Each supervisor profile is defined by a list of keywords (classification terms) used in the DiVA theses that have been supervised by the researcher in question to describe their area of expertise.

Data

The data set consists of 2,216 records of student theses published between 2010 and 2017. One issue with data entered during a long interval like this is that the quality of the records varies. In the case of our sample data this certainly is true. In 2015 the BTH repository joined the national Swedish consortia DiVA-system. At around the same time the library started supplying records to a national portal “SwePub” intended to become a national tool and a resource for the collection of publication data for bibliometric analyzes and data processing (SwePub). The DiVA migration and having SwePub harvesting our records meant that the library needed to focus much more attention on the quality of the records. The amount of metadata attached to every record increased.

Records entered by students or staff themselves usually have quality issues. In our case we have quality issues with records entered before the second half of 2015. Beginning in the fall of 2015 records are always quality checked by staff from the library or the university departments.

Another data issue is knowing what expert in the used data set still is affiliated with BTH. In a data set covering eight years many supervisors have come and gone. In our
case we needed a big data set to create and validate a prototype system so the issues discussed above are not really a problem. But if you want to put an Expertise Recommender System into everyday use issues like these must be paid close attention to.

**Ontology Description**

We have defined our ontology as a formalization of concepts and relations between them. The domain concepts are described by classes and the relations between them is defined as “isSubclassOf”. Main focus in our ontology are classes and their subclasses. The model in this study is created using National Subject Categories in DiVA (Linköping University Library) that is standard for the Swedish listing of research subjects 2011 and classification terms from extracted data. National Subject Categories in DiVA are separated into three levels. Namely, the main categories (the highest level) are Agricultural Sciences, Engineering and Technology, Humanities, Medical and Health Sciences, Natural Sciences, and Social Science. These main categories are located at the first level of the ontology. The others two levels are described by the sub-subject areas that belong to each of the main categories. The second level has 26 subject categories and the third level consists of 257 categories, respectively. From the extracted data we have taken the subject terms (specific domain keywords) and added them to the structure of ontology model as classes. In that way, the created ontology tree has a four level depth.

**Metrics**

Silhouette Index (Rousseeuw) is a cluster validity index that is used to judge the quality of any clustering solution $C = \{C_1, C_2, \ldots, C_k\}$. Suppose that the considered data set contains the attribute vectors of $m$ objects. Then the Silhouette Index is defined as

$$s(C) = \frac{1}{m} \sum_{i=1}^{m} \frac{(b_i - a_i)}{\max\{a_i, b_i\}}$$

where $a_i$ represents the average distance of object $i$ to the other objects of the cluster to which the object is assigned, and $b_i$ represents the minimum of the average distances of object $i$ to object of the other clusters. The values of Silhouette Index vary from -1 to 1.
Results

As it was mentioned above, the test data is downloaded from BTH DiVA installation, concerning thesis supervision of researchers affiliated with BTH. The extracted data has been used to construct the supervisor profiles and ontology model. Each supervisor profile is defined by a list of keywords (domain-specific terms). After resolving the problem with ambiguity, the set of supervisors has been reduced to 375 expert profiles.

We have studied two different experiment scenarios. In the first scenario, the user is supposed to provide an example supervisor and the developed recommender system will return a list of experts (supervisors) with close (similar) expertise. In this scenario, the system initially calculates the expertise similarity scores between the given supervisor and all other supervisors presented in the system by using the expertise similarity formula described in section “Expertise Similarity”. Then the system returns a ranked list of similar experts who are recommended subject thesis supervisors. For instance, we have tested the system by using an example supervisor whose expertise can be described by the following subject terms: database, performance, usability, web server, cloud computing and Amazon web services. The system has returned a list of the supervisors ranked with respect to the similarity of their expert profiles to the profile of the example supervisor. In table 1, we list the ten top ranked supervisors who have been recommended by the system. The calculated expertise similarity scores between the example supervisor and the all other supervisors presented in the system are in the range between 0.046 and 0.77. The expertise of each top recommended supervisor is described by the keywords found in their expert profile (the second column in table 1). As one can notice the all top ranked supervisors have expertise that is overlapped to different extent with the competence of the given example supervisor. For example, the expertise of supervisor 1 covers usability subject from the example expert profile while supervisor 8 has competence overlap in two subjects: database and performance.
Table 1.

<table>
<thead>
<tr>
<th>Expert</th>
<th>Keywords</th>
<th>Expertise similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>user experience, usability</td>
<td>0.770</td>
</tr>
<tr>
<td>2</td>
<td>privacy, security, cloud computing</td>
<td>0.763</td>
</tr>
<tr>
<td>3</td>
<td>cloud computing, security metrics, security threats, security measurement frameworks</td>
<td>0.760</td>
</tr>
<tr>
<td>4</td>
<td>procedural city generation, perlin noise, performance, game content</td>
<td>0.760</td>
</tr>
<tr>
<td>5</td>
<td>machine learning, parallel computing, multiprocessor, performance</td>
<td>0.760</td>
</tr>
<tr>
<td>6</td>
<td>mobile, power, consumption, android, native, web, enterprise service bus, performance, framework</td>
<td>0.754</td>
</tr>
<tr>
<td>7</td>
<td>mongodb, couchdb, python, pymongo, couchdb-python, nosql, document database, json, dbms, database</td>
<td>0.754</td>
</tr>
<tr>
<td>8</td>
<td>compression, sms, arithmetic, lambda, huffman, lzw, lz77, lz78, fristående kurs, voltdb, mysql, databases, main-memory database, performance</td>
<td>0.751</td>
</tr>
<tr>
<td>9</td>
<td>non-functional search-based software testing, non-functional system properties, search-based software testing, meta-heuristic techniques, performance testing, load testing, load patterns</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>digital multimedia broadcasting, mpeg-2 standard, mpeg-4 standard, video transport stream</td>
<td>0.75</td>
</tr>
</tbody>
</table>
In the second scenario, the set of supervisors is clustered into groups of experts with similar expertise by applying $k$-means partitioning algorithm. The optimal number of clusters is determined by clustering the set of supervisors applying $k$-means for different $k$ and evaluating the obtained solutions by the Silhouette Index (SI). We have applied $k$-means for $k$ between 2 and 257 (the maximum domain-specific classes in ontology). The selected optimal value for $k$ is 16 and the corresponding average SI score given on the clustering solution generated for $k = 16$ is 0.38. In this scenario, in order to select the right individuals for a specified thesis subject the user may restrict their considerations only to those supervisors who are within the cluster that is identical with (or at least most similar to) the given subject. The experts in the selected cluster can be ranked with respect to the similarity of their expert profiles to the specified subject. For example, if we need a supervisor with expertise in database and parallel computing we will use cluster 0 (see table 2), in which all experts have competence either in one or in both fields.

**Table 2.**

<table>
<thead>
<tr>
<th>Experts</th>
<th>Cluster</th>
<th>Description of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>3, 4, 7, 8</td>
<td>0</td>
<td>usability; tessellation; android; security threats; main-memory database; database; distributed databases; parallel computing; security; data mining, etc.</td>
</tr>
<tr>
<td>2, 5, 6, 10</td>
<td>3</td>
<td>usability; data mining; performance monitoring; systematic review; video streaming; parallel computing; mpeg-2 standard; mpeg-2 standard; nosql database; machine learning; cloud computing, etc.</td>
</tr>
<tr>
<td>1, 9</td>
<td>15</td>
<td>usability; quality of experience; urban design; systematic literature review, etc.</td>
</tr>
</tbody>
</table>

We have checked how the 10 top ranked supervisors returned by the system in the first experiment scenario has been distributed between the generated clusters. The obtained grouping is presented in table 2. As we can see the supervisors who have been grouped together have a similar area of competence. For instance, supervisors 3, 4, 7 and 8 are distributed in one and the same cluster (cluster 0). The main
expertise of supervisors 3 and 4 is in cloud computing while supervisors 7 and 8 both have competence in database. The supervisors 2, 5, 6 and 10 are grouped in the cluster 3. For example, the supervisors 5 and 6 have expertise in performance and the others are in cloud computing and video streaming. However, supervisors 1 and 9 are in a different cluster (cluster 15 in table 2), because as we can see in table 1 they have very unique expertise in comparison with the others.

It is interesting to notice that the generated clustering solution partitions the supervisors into 16 disjoint clusters (subject categories), i.e. each supervisor belongs to only one cluster. However, as it can be seen in table 2 there are subject terms that participate in the description of more than one subject category, e.g., parallel computing is used for clusters 0 and 3 (the third column in table 2), i.e. the obtained subject categories can have an overlap on some topics.

Conclusion and Future Work

The main contributions of this work are:

• a prototype system for ranking potential thesis supervisor candidates based on their previous supervision experience and expertise retrieval approaches;

• an evaluation of the prototype with real-world data extracted from the BTH instance of the institutional repository system DiVA.

The proposed system is based on online information, can be updated regularly, and uses standardized subject vocabulary, i.e., DiVA classification terms associated with each thesis. It can be implemented by any university that uses an institutional repository with a controlled vocabulary including keyword fields in the records, i.e., it is general.

For future work, we aim to pursue further evaluation and validation of the developed expert recommended system on bigger datasets and also on data extracted from other DiVA university repositories. Our future plans also involve using additional information about the supervisors’ expertise, e.g., extracting information about the supervisors’ research publications, in order to build richer expert profiles and ontology model.
References


The Public Knowledge Project
Reflections and Directions After Its First Two Decades

Juan Pablo Alperin
Simon Fraser University
juan@alperin.ca

John Willinsky
Stanford University
john.willinsky@stanford.edu

Brian Owen
Simon Fraser University
brian_owen@sfu.ca

James MacGregor
Simon Fraser University
jmacgreg@gmail.com

Alec Smecher
Simon Fraser University
alec@smecher.bc.ca

Kevin Stranack
Simon Fraser University
kstranac@sfu.ca

Abstract: After more than 20 years in operation, the Public Knowledge Project (PKP) faces a sizeable responsibility to the more than 9,000 journals that rely on its open source publishing software. In light of this responsibility, in 2017 PKP felt compelled to undertake a large-scale strategic review and community consultation to better understand the maturing of, and challenges faced by, an open source software project seeking to sustain increased global access to research and scholarship. This paper presents the result of the consultation carried out by the consulting firm BlueSky to BluePrint, along with PKP’s response to those findings. In doing so, it identifies the three pillars of PKP’s work: Open Source Software as well as Research, Education, and Advocacy. In what may serve as a lesson for other open source projects, the paper concludes with a path forward for PKP.

Keywords: scholarly publishing, open access, open sources software, public knowledge, scholarly communication.
Introduction

As the Public Knowledge Project (PKP) enters its third decade, it faces the responsibilities of supporting the more than 9,000 journals using its software which are dependent on PKP continuing to develop the code. In the fall of 2017, PKP, with the support of the Arnold Foundation, contracted the consulting services of BlueSky to Blueprint, with its principal Nancy Maron embarking on an exploration of PKP’s standing and prospects among a sample of those involved in scholarly publishing, including current, former, and potential users of its software (Maron).

Getting unvarnished feedback from the community, through a third party, has given us the opportunity to see ourselves through the lens of others. This view has enabled us to reflect on our values, our activities, and our operations. We are certain of several things: we have at heart our mission to provide free open source software to anyone, anywhere, as a means of lowering barriers to creating, presenting, and sharing scholarly content. We are an academic-led initiative and are firmly a part of the community we serve. We will never be acquired by a commercial entity. In order to continue to provide best-in-class service to our thousands of users worldwide, our not-for-profit, academic-led initiative needs to operate in a way that makes best use of business strategy to ensure that our products and services are competitive with offers of all types, and that our internal operations are efficient and productive. Our community of users is everything to us, and we will find more ways to actively support, educate, and engage them in the years ahead.

At the same time, it has become clear that we need to find new ways to better support our work. To this end, we present a summary of our interpretation of the findings, both the positive and negative, along with the actions we intend to take to improve how we operate. This is surely not the only way to interpret the findings report, but it reflects what we, the PKP Team, have learned about ourselves, and the vision that we have for the organization. We welcome additional input, both in response to this document and to the report that gave rise to it, as we move forward on implementing.

Our Three Pillars

Through this process, a picture emerged of three areas of strength for PKP. These three distinct but interrelated activities are at the heart of PKP. Collectively, they define PKP and they explain how PKP serves its larger goal of making knowledge public:
• Our open source software mission drives everything we do. We remain committed to developing excellent software and to making it freely available to all. While OJS is our flagship software, we will continue to be involved in other areas of scholarly communications, like our monograph software (OMP) and our Typesetting stack (OTS). Being an open source project is important to us because it builds capacity around the world by encouraging a “do-it-yourself” approach; it allows PKP to benefit from outside contributions, such as translations and code; it supports more diverse use-cases, allowing users to customize their installations to suit their needs; and it demonstrates that openness is at the core of PKP.

• Our research, education and advocacy are a core part of who we are. We are deeply embedded in the academic, library, and technical communities that make up higher education, and seek to both build on and contribute to our collective understanding of the ways in which that knowledge is created, circulated, and used, under the belief that the public has a right to access knowledge. To this end, we research a broad range of issues around scholarly communications, innovate and experiment with new technologies, and work to educate and be actively involved with the academic and scholarly communication communities. Doing so helps us achieve our goals because:
  a. the research builds a better understanding of the issues around scholarly communication;
  b. the research and our active involvement at events keeps the PKP team embedded within the academic and scholarly communication communities we seek to affect;
  c. our training materials and workshops help raise awareness and improve the quality of scholarly publishing;
  d. the experimentation and innovation with technologies improves the software;
  e. and being a voice in the community helps directly influence decision making and discussions.

• Our PKP Publishing Services (PKP|PS) are a key way in which we support those in the community who prefer not to host software themselves, as well as to those seeking a community-based not-for-profit alternative for their publishing needs. Importantly, providing publishing services supports PKP because they:
  a. provide revenue back to the organization to support the software, research, education, and advocacy work;
  b. and gives us the constant feedback we need for our software to remain competitive with large commercial offerings (e.g., ScholarOne).

After listening to the community, we have come to believe, more strongly than ever, that it is essential for PKP to continue to operate in all three areas. However, it has also become evident that each undertaking of PKP can no longer be done...
in isolation from the rest. PKP is committed to bringing these activities together in ways that reinforce each other and are philosophically coherent and consistent with the principles that underlie the open access community.

Our core take-away from this report is that PKP should strengthen the ties among its activities, and actively work to communicate more about those linkages to the community. Doing so will not only help others to assess the value of PKP, it will also provide a stronger technological, social, and financial foundation that will enable PKP to continue to have an impact that is disproportionately large for its modest budget and team size.

To strengthen the ties between our activities, we realize we also need to build out each area. We have taken stock of the feedback we received from the experts and stakeholders we consulted, and of our own sense of where our strengths and weaknesses are. There is much to be proud of in each realm, but also many opportunities for improvement. We summarize some of the key findings and our proposed approach to move forward below.

**Open Source Software**

*The good news:* OJS is the reason that a large number of journals are able to publish. Individuals and organizations around the world are able to set up an instance of OJS and operate a journal because it is free and open source, and because it has minimal server requirements. There are over 9,000 journals actively using OJS (publishing at least 10 articles a year, and an average of 46 articles per year). Many find OJS helpful for the ways it can be configured and the range of publications it makes possible. Those who have used or seen OJS3 are very pleased with the more modern design and upgraded features.

*The less-good news:* some are frustrated by what they perceive to be an unnecessarily slow pace of change. Many perceive OJS as being outdated and rigid, both in its look and its workflow, based largely on their experience with OJS2. Some have stated they do not feel PKP functions as fully “open source” as it could and should be more actively seeking code contributions from others. Some feel OJS is “too heavy” and requires using or dealing with more workflow elements than are needed. There is little awareness and lower levels of adoption of the other PKP software (e.g., OMP and OCS).
Next steps: we plan to actively encourage open source contributions from the community. Doing so will increase the speed at which we can deliver new features and respond to the needs of the community, like we have seen with OJS being translated by volunteers into dozens of languages. To do this, we will:

- find ways to call attention to the many contributions our community continues to make to our code and to our documentation;
- review and improve documentation for the developer community;
- revise and redesign the online spaces (such as the PKP Forum and Github), so that it can serve as an obvious way to encourage recommendations for new features, have the community vote on them, and track progress as new efforts progress;
- specify discrete and actionable work “packages” or assignments for outside contributors;
- dedicate PKP developer time to facilitate code contributions from those that express interest in customizing their own installations;
- host more developer-centric events (e.g., code sprints);
- decrease the code complexity by incorporating third party software and updating/streamlining code to be more transparent.

We also plan to make our software easier to use and adopt. We are certain that OJS delivers value (9,000 journals cannot be wrong!). But still, people have issues with how it looks, operates, and what it can do. To address these issues we will:

- develop tools, documentation and service packages to make it easier for existing OJS2 users to migrate to OJS3, so they can enjoy its full benefits;
- review the OJS interfaces and workflow to see if there are ways of offering a more streamlined experience for those not requiring all the functionality it offers;
- communicating more effectively the modular and customizable aspects of OJS3, by offering visuals, documentation and other guidance on ways it can be used in part or in whole;
- invite more community input at the vision, planning and priorities stages;
- make it easier for users to see how OJS can function in harmony with other popular software and infrastructure, including open source repositories (e.g., DSpace, Islandora, Samvera) and centralized services (e.g., Crossref, ORCID).
Research, Education, and Advocacy

The good news: there is a high degree of trust, admiration and respect for PKP. The vast majority of those we consulted held PKP in high regard, and were appreciative of the work that we do. We are actively publishing and presenting at academic venues around the world, with both invited and peer-reviewed activities; our research is widely read and used to inform the collective understanding of scholarly communication (e.g., Alperin and Rozemblum; Alperin, Gomez and Haustein; Willinsky; Willinsky and Alperin). We are a strong voice for scholarly publishing in the academic community, with PKP team members serving on editorial boards and advisory groups of important initiatives such as SPARC, Crossref, Open Library of Humanities, HIR-MEOS, ScholarlyHub, DOAJ, Redalyc and the Library Publishing Coalition. Our educational offerings have made a significant contribution towards professionalizing the work of new editors and emerging publisher. We have successfully moved some experimental services, like the PKP Index and the PKP Preservation Network into production offerings, open to all journals that use OJS. We are developing new community-based publishing models, whether nationally with Coalition Publi.ca in Canada or within disciplines, such as Libraria in anthropology and working with similar initiatives around the world, e.g. Journal.fi (Finland).

The less-good news: there is a very limited awareness of PKP’s activities outside of our developing OJS. Some believe that PKP’s software needs to be more innovative and keep up with the times. Some have expressed frustration with their efforts to collaborate with PKP, following promising conversations with PKP team members at events and meetings. Our capacity to deeply interact with a very large and geographically diverse community has been limited to attending and putting on conferences and to the PKP Forum. PKP’s educational content requires regular updating and expansion to better serve the community, and more effort is needed to raise awareness of activities in this area. Prototypes, like article-level metrics (Alperin et al.) and automated XML typesetting (Garnett et al.) have not yet made it into production, despite their importance to the community.

Next steps: PKP’s strong reputation gives us the opportunity to play a leadership role in supporting the academic community, but to do so, we need to:

- reinvest in community outreach, as a formal activity, offering more ways for us to stay in touch with our users, more ways to support them and help them to help each other;
- re-evaluate our educational activities to foster a shared understanding of the values and practices between PKP and the community;
• overhaul our web presence, documentation and other materials to be more inviting of input and dialogue with the community, as well as to better showcase the work and activities PKP is already involved with;

• expand the reach and diversity of PKP’s community-based committees with a particular emphasis on guiding PKP’s work across all areas;

• build on the global reach of PKP by hosting or helping to organize events with user groups around the world.

We also need to formalize processes for research and innovation activities to improve our offerings. We have been successful as a research group, but our research does not always make its way into our offerings. Similarly, we have had mixed success in moving experimental services into production. To strengthen the ties between our research, software, and services, we will:

• develop and seek funding for a research agenda that addresses pressing needs of the academic community, while simultaneously providing useful intelligence to further PKP’s development;

• create something akin to a “PKP Labs,” as a means of identifying, building and testing innovative projects, supported by internal and external funds.

**PKP Publishing Services**

*The good news:* we have created a publishing services offering (PKP|PS) that has continued to grow since it was started in 2007, to a point where today it supports approximately 50% of PKP’s annual operating expense. Our clients report a high level of satisfaction with the degree to which we are able to customize our software to suit their needs. We serve a diverse set of clients, including everything from small, independent publishers with one journal, to major society publishers with many high-profile journals.

*The less-good news:* there is limited awareness of PKP|PS and the services it offers. There are many competitive services available that host OJS; some are in the academic space, some are commercial offerings. Most do not contribute anything back to PKP. We know the degree of support top-tier publishers require and to expand our offering, we will need to substantially increase our staff support.

*Next steps:* our publishing services have grown to make up over half of our operating revenues, and they have done so without any targeted promotion. Although some clients seem to find us through word of mouth or through our web page, too many
potential clients are unaware of our services. We will promote and grow our services by:

- having members of the PKP|PS team participate in industry conferences, as well as academic events, for the purposes of promotion, hiring new staff if necessary;
- developing and executing a full marketing strategy that is consistent with PKP’s mission and values, while simultaneously extolling the quality of PKP|PS.

Our current clients include single-journal publishers and multi-journal publishers, commercial and non-profit organizations. However, we are aware that certain types of publishers do not see us as an obvious choice, so we have decided to:

- develop service packages that can be customized to address the needs of a certain type of user/potential client;
- build partnerships with other service providers, or expand our own services, to offer additional options to all clients, including professional publishing and editorial support services, theming, and customization;
- present an array of service options and packages in ways that make it easier for the wide range of people who make purchase decisions.

The publishing services team already runs a successful operation that generates surplus revenue, but more can be done to improve its efficiency and increase the surplus that supports other PKP activities. We will improve operations by:

- investing in building tools that the PKP|PS team can use to further automate installing, upgrading and monitoring journals;
- offering an option to use a shared installation (i.e., cloud-based) for journals that do not require customizations to minimize the maintenance overhead;
- revisiting our pricing model, to make it more appealing for large institutions and publishers to host a large number of journals with us.

A Way Ahead

After two decades of being an academic-led “project” within the university context, the PKP organization has successfully matured into a reliable, secure, and evolving organization with a range of related goals and activities, and a globally recognized contribution to make to scholarly communication and public knowledge. We have formed a dedicated team, many members of which have now been with PKP for well over a decade, and all of whom are experts in our software and in supporting our com-
munity around the world in various ways. That said, our internal operations are ready for a new phase of growth and structure. In order for the plans outlined in this report to take flight, we understand we will need to make internal changes by, for example, supplementing existing staff with those who can bring additional skills to our team.

Organizational Structure and Governance

Our organization currently includes an internal management team and a few advisory groups. We are based at Simon Fraser University, in Canada, with roots in both SFU Library and the School of Publishing. We plan to keep these elements in place going forward, but are committed to revisiting the governance structure, the make-up and mandate of the advisory groups, and the mechanisms we use to gather ongoing feedback and direction from our community.

Internally, we have already begun to see how thinking about PKP’s work around the three pillars outlined above can be reflected in the financial management of the organization. As we seek to strengthen each pillar of activity, we are ensuring that each area operates under its own internal logic, with specific goals that can be tracked. By having these be reflected in the financial management, we are ensuring each area is adequately resourced, and is not made to suffer at the expense of the others. Conversely, we are also making explicit the direct and indirect supports that each provides the others, so that when resources need to shift from one priority to another, it is done so with a clear rationale and well-defined parameters.

Externally, we recognize the need to restructure our decision making to have greater involvement from the community. We see potential in a membership model, with members having voting rights to help us determine key elements on our development roadmap and the project’s overall direction. The actions outlined above are also largely oriented to capturing more community feedback, regardless of membership, and so we intend to formally bring that feedback into decision-making going forward.

Internal and External Communication

One of the main findings of this study was that we need to better communicate how PKP contributes to open scholarly publishing. Communicating who we are and what we do is built into the proposed actions of each of the pillars above, but we feel it is important to highlight separately, as it is essential that PKP also present the unified picture of how the pieces fit together. To do so we need to develop a formal
communication strategy that can be used within the organization to ensure we have a shared understanding of our goals and activities, and externally so that the scholarly publishing and academic communities can fully benefit from and appreciate the work that we do.

This communication strategy needs to be bidirectional, and will require one or more additional team members who can actively do outreach at major community events, participate in important meetings, and regularly solicit frank feedback from stakeholder groups on directions and details (like this project did).

**Sources of support**

PKP has just concluded a five-year sustainability strategy that saw our operating budget grow substantially, and which has put us on much firming footing going forward. This growth has come in large part from the PKP’s Publishing Services (PKP|PS), but continues to be supplemented by our sustainers program and by significant grant funding. We have also benefited from community contributions through things like translation of our software and some code contributions. However, we are convinced that we are not maximizing the potential in any of these areas.

The actions outlined above will go a long way to increase the in-kind support we get through community contributions. Our inability to grow the pool of contributors substantially over the years has limited our growth and we are certain that by having dedicated personnel who will be in charge of outreach, documentation, and development (in the specific ways outlined above), we will translate into growth the contributions we receive in various forms (e.g., translations, code, hosting PKP events, etc.). The value of these contributions, we expect, will far outweigh the efforts and costs necessary to generate them.

However, to grow the team with people who are capable of encouraging and supporting more community contributions, we need to have the necessary financial resources. We see potential in both models recommended: a membership model that can bring both financial and in-kind contributions (as well as feedback and direction, as noted above), and through the growth of PKP|PS.

To grow our membership model, we are reviewing ways to make participation more attractive to a wide number of our users. Some may want to affiliate with PKP for benefits like discounted registration for PKP conferences and workshops. Others—those with deeper ties to PKP—may want to support the organization financially. We
will review our key relationships, and seek to establish closer ties to organizations that are built upon our software, to see how their success can also help fuel our ongoing support of their work. This work will itself require dedicated people to build relationships, but after an initial investment, the financial contributions of members should make this position self-sustaining.

To grow our PKP Publishing Services, we need to first ensure that people become aware of the services we offer, and subsequently review our offerings to provide the level of support that even the largest and most sophisticated publishers expect. We have been fortunate until now to not have to actively promote our services, but we realize that this has limited our capacity to grow. Growing PKP|PS will require bringing on, at the very least, a dedicated business development manager and customer service support specialists. The business development manager will be seeking clients from organizations that support multiple journals, and will actively work with university-based publishers to introduce them to our offerings, as well as seeking out other segments for growth. Like with the personnel necessary to grow the membership model, the costs associated with these positions should be self-sustaining through growth in the business after an initial investment. Moreover, the outreach that helps grow the business will also be a key element in ensuring that those who use OJS and our other software feel fully a part of the broader PKP community.

Being an academic-led project, we will also continue to pursue grants as a way to support the research and innovation of PKP, but hope to use the new organizational structure to remove the reliance on such grants for core operations. Grants, unless specifically intended for operations, should be focused on the research activities of PKP, with the intended benefit to other areas, but without becoming a source of direct support for non-research activities. Taken together, these efforts to both raise revenues, streamline processes, and expand the membership base of the organization will permit the PKP to quickly expand the speed of development, and to remain in close touch with the very people it was created to serve.

**Conclusion**

This study served as an inspirational moment for our team in its commitment to supporting greater access to research and scholarship into a third decade. It has helped us gain insight into what distinguishes PKP’s contribution and what people value about our approach. We are more certain than ever that PKP should sustain its efforts in the three designated areas, while realizing that there is always more to do in fulfilling the resulting responsibilities:
to communicate the range and complexity of our undertakings;

• to be responsive and attentive to academic, scholarly publishing and open sources communities;

• and to collaborate with a growing number of organizations to advance the quality and openness of research and scholarship.

After a thorough consultation with members of the scholarly publishing community, Nancy Maron concluded that PKP has strengths to build on in undertaking the change that is needed to better serve its community:

PKP has developed a deep and loyal global following over its nearly twenty years of operation. It is poised today to squarely addresses its challenges and leverage its strengths into a dynamic plan for the future. The next steps will almost certainly involve a good deal of change, and this may not be easy. But publishers of all types in countries around the globe, continue to seek ways to create, consume, publish and share information, and the Public Knowledge Project, through its mission, its software, and the support of its community, is in a position to help them as needs continue to evolve. (Maron)

Our plan is to use what we have learned from the Community Consultation that Nancy Maron conducted to not only catch up with the growth in the global use of our software, in terms of our publishing services and support, but to push the thinking about ways of improving scholarly communication even further, in pursuit of such goals as universal open access and community control of scholarly publishing infrastructure.

Going forward, we intend to facilitate increased community contributions and input throughout the areas in which we work; to strive for more internal coherence and cross-over among what we now think of as PKP’s three pillars; to give the library community and granting agencies more reasons to sponsor both our publishing innovations and our software upgrade cycles. More specifically on this last point, we see exciting opportunities ahead for developing editorial efficiencies for improved research quality, reader enhancements through design improvements, stronger integration of infrastructure for improved scholarly communication at every phase of the research process. We can also see now that our aim is to build a publishing service that offers both a level of professionalism (in its responsive service focus) and a commitment to being of, as well as for, the academic community of which it remains proud to be an integral part.
References


Beyond the Dichotomy between Natural and Knowledge Commons

Reflections on the IAD Framework from the Ubatuba Open Science Project

Sarita Albagli
Brazilian Institute of Information in Science and Technology, IBICT – Brazil
sarita.albagli@gmail.com

Anne Clinio
Liinc – Interdisciplinary Laboratory of Information and Knowledge Studies, Brazil
anneclinio@gmail.com

Henrique Z.M. Parra
Unifesp–Federal University of São Paulo, Brazil
henrique.parra@unifesp.br

Felipe Fonseca
Ubalab, Brazil
felipefonseca@gmail.com

Abstract: The paper presents a critical analysis of the possibilities and limits of the Institutional Analysis and Development (IAD) framework, proposed by Elinor Ostrom and team, specially addressing the mutual relations between natural and knowledge commons. It results from an action research project on the role of open science (OS) in development, carried out in the municipality of Ubatuba, on the North Coast of the State of São Paulo, Brazil, in 2015-2017. The work involved: systematizing the literature on the IAD framework; mapping and selecting literature representative of other theoretical and conceptual approaches; critically using and adapting the framework to the case studied. The project provided the opportunity to observe how these dynamics take place in a relatively small-scale (while heavily interconnected) context. While the IAD framework helped us to analyze the institutional, political, and governance issues affecting knowledge production and circulation, we observed the higher complexity of our action arena, shedding light on the fact that natural and knowledge commons are the two dimensions of the same “commoning” process.
Keywords: Commons, Open Science, Institutional Analysis and Development, Ubatuba, Brazil.

This work was funded by OCSDNet/IDRC/UKAid, Fundação Carlos Chagas de Apoio à Pesquisa do Estado do Rio de Janeiro (Faperj) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We also thank the contributions of the reviewers.

Introduction

The paper presents a critical analysis of the possibilities and limits of the Institutional Analysis and Development (IAD) framework, proposed by Elinor Ostrom and researchers from Indiana School, specially addressing the mutual relations between natural and knowledge commons. It is based on the results of an action-research project on the role of open science (OS) in development, carried out in 2015–2017, as part of the Open and Collaborative Science in Development Network – OCSDNet. Focusing on the institutional, political, and governance issues affecting knowledge production and circulation, the project provided the opportunity to observe how these dynamics take place in a relatively small-scale (while heavily interconnected) context—the municipality of Ubatuba, on the North Coast of the State of São Paulo, Brazil. Our study produced rich empirical and theoretical material for analysis, offering possibilities for critical reflection as well as social learning relevant to other territorial and social contexts (Albagli et al.).

Ubatuba is located in the Atlantic Rain Forest region, a strategic and vulnerable environmental area, with a high level of endangered socio-biodiversity, and a focus of intense scientific research. Ubatuba’s key development challenges are related to how to conciliate:

• its rich and strategic natural, cultural and knowledge commons;
• the necessity to provide access to local populations to social and economic benefits derived from the use of that wealth from a sustainable development perspective;
• political empowerment of local communities in a context of inequality of access to institutional deliberation processes;
• and the contributions that information and knowledge may make for these processes.

While most of the Ubatuba’s territory (around 80%) is located within the protected area of the State Park of Serra do Mar (PESM), its economy is based on seasonal and predatory tourism that encourages real estate speculation, as well as, more recently, oil exploration boosted since the pre-salt discoveries. These aspects characterize a highly contentious action arena regarding its natural and immaterial commons.

This paper presents a part of the research and it involved the following steps:

- Systematizing the literature on the IAD framework, in order to understand its rationale and consider its possible uses (and limits) in our case study. We were particularly interested in understanding how this framework expanded to include knowledge commons as part of its analysis.

- Mapping and selecting literature representative of other theoretical approaches to the concept of common(s) (Hardin; Bollier; Hardt and Negri; Dardot and Laval; among others), observing their convergence and divergence with Ostrom’s perspective.

- Developing a two-way exercise. On the one hand, we mobilized aspects of the IAD framework as a tool kit to help us select and organize relevant information, to characterize our “action arena” and to define an “action situation”, focusing on the local socio-institutional context, key actors and their (cooperative or conflictive) relationships. We expected this approach would be helpful to analyze our case, because it opposes a path dependence perspective, giving place to future alternative scenarios and it could be used to analyze dynamic and changing situations. On the other hand, we confronted the IAD framework with our empirical research results, also considering other interpretative approaches identified in previous steps.

At the end we observed that the relevance of IAD framework lies in the fact that the diffusion and adoption of open science is closely related to institutional issues (both formal and informal) that affect the open and collaborative nature of knowledge production and circulation. On the other hand, those issues are inextricably invested with conflicts and power relations over natural and immaterial commons. In this sense, it has also highlighted the mutual and contradictory relations between the new infrastructures and the vulnerability/robustness of information and knowledge commons, which requires going beyond the access paradigm.
The IAD Framework and the Commons

The IAD framework was developed by Elinor Ostrom and other researchers of the Indiana School, based on extensive empirical research that demonstrated that a community can self-organize to “successfully” use and manage a common pool resource (CPR) (Ostrom, Governing the Commons). In other words, they argued that local and self-organized populations can economically exploit a CPR in a sustainable way for long periods of time. The IAD framework was first built on research on urban public goods, and it was further developed based on the work on the formal and informal rules that positively or negatively affect the sustainable management of natural CPRs (such as groundwater basins, irrigation systems, grazing systems, and forests). Ostrom’s team main question was:

. . . how a group of principals who are in an interdependent situation can organize and govern themselves to obtain continuing joint benefits when all face temptations to freeride, shirk, or otherwise act opportunistically? (Ostrom, Governing the Commons 29)

The results from the case studies helped them to question widely accepted theories—such as “The Tragedy of Commons” (Hardin), “The Prisoner’s Dilemma” (Dawes, Formal Models; Dawes, The Commons”) and “The Logic of Collective Action” (Olson)—for whom individuals necessarily develop opportunistic behavior towards the maximum exploitation of common resources, putting individual profit above all, and disregarding the collective losses of overexploitation. According to those theories, predatory behavior is an inherent feature of collective management of common resources, which necessarily leads to their ruin. This would justify the prescription of either the privatization of the commons or the imposition of rules by the State. In all cases, those theories envisaged the necessity of an external authority to supervise the use of common resources either by limiting their access or by applying sanctions to those who violate the rules established to ensure long-term sustainability and productivity.

2. In 2009, Ostrom shared the Nobel Memorial Prize in Economic Sciences with Oliver E. Williamson for “her analysis of economic governance, especially the commons” (The Royal Swedish Academy of Sciences).
3. For Ostrom, the notion of “successful” management of common goods refers to the twofold objective of (1) avoiding overexploitation and exhaustion of resources, and (2) generating quality of life, understood here as the effective people’s participation in the management of the common goods through principles of coexistence shared by all involved.
4. “. . . it refers to a natural or man-made resource system that is sufficiently large as to make it costly (but not impossible) to exclude potential beneficiaries from obtaining benefits from its use.” (Ostrom, Governing the Commons 30).
Commons were later defined as a general term referring to “a shared resource that is vulnerable to social dilemmas” (Ostrom and Hess 13), meaning: high vulnerability to subtraction (also referred to as rivalry, when the use of something by someone prevents its use by another one) and difficulty to exclude free riders (opportunistic behavior). Ostrom and Hess did not differentiate common (singular) and commons (plural). For them, “Commons is an awkward word in the English language. The same word is used for both the singular and plural forms.” (Ostrom and Hess 21). Other authors would argue that the distinction between these two terms is a significant point, as we will discuss later.

Ostrom and Hess focused on the institutional dimension: “the rules, decisions, and behaviors people make in groups in relation to their shared resource” (10). A set of eight “design principles” of institutional robustness in (un)successfully managing common-pool resources were pointed out (Ostrom, Governing the Commons), provided that they should not be seen in a prescriptive way, as models, but rather as “insightful findings in the analysis of small, homogeneous systems” (Ostrom and Hess 7). Their institutionalist approach emphasized rules-in-use (those practiced by actors) and “invisible” institutions (widely recognized sets of rules-in-use), as key aspects at times of institutional change. In this sense, institutions were conceived as:

...formal and informal rules that are understood and used by a community...[They] are not automatically what is written in formal rules. They are rules that establish the working “do’s and don’ts” for the individuals in the situation that a scholar wishes to analyze and explain. (Ostrom and Hess 42)

Ostrom was particularly interested in developing a “microsituational level” of analysis and “more configural approaches”, based on empirical work in order to confront “an immense diversity of situations in which humans interact”, and to stress the importance of fitting policy prescription and institutional rules to specific social-ecological settings.

‘One-size-fits-all’ policies are not effective. (Ostrom, Beyond Markets 409)

The authors also highlighted the role of informed and communicative patterns of interaction within the community as ways to develop a common language and a collective understanding of the use of common resources, elaborating norms on rights and duties in resource management and evaluating the cost-benefits of agreed rules.

5. Vincent and Elinor Ostrom’s previous work had typified four kinds of goods—public goods, toll or club goods, common-pool resources, and private goods, classified according to their level of exclusion (easy, difficult) and subtractability (low, high).
With adequate information [participants] may develop increasing trust so that the situation can lead to productive outcomes. (Ostrom and Hess 59)

The IAD methodology was conceived as a “multitier conceptual map” (Ostrom, *Understanding Institutional Diversity*) and a “metatheoretical language”, designed:

... to enable scholars to analyze systems that are composed of a cluster of variables, each of which can then be unpacked multiple times depending on the question of immediate interest.” (Ostrom, *Beyond Markets* 414)

It is organized in three clusters of variables schematically represented by figure 1. Depending on the research question, it is possible to privilege one of these clusters as the starting point of the analysis.

**Figure 1. Institutional Analysis and Development Framework**

The action arena refers to “the social space where individuals interact, exchange goods and services, solve problems, dominate one another or fight (among the many things that individuals do in action arenas)” (Ostrom, “Institutional Analysis and Development”). Action arenas include one or more action situations and the participants in those situations (Ostrom, *Understanding Institutional Diversity*). The action situation,
enables an analyst to isolate the immediate structure affecting a process of interest to the analyst for the purpose of explaining regularities in human actions and results, and potentially to reform them. (Ostrom, “Institutional Analysis and Development” 268)

Action arenas and action situations are considered to be at the core of the IAD framework and they are affected by a set of broadest categories of “external factors” (biophysical characteristics, attributes of the community, and rules in use).

Traditional and New Commons

Ostrom’s initial focus on natural commons (Ostrom, Governing the Commons) was further expanded to the analysis of information and knowledge as “new commons” (Ostrom and Hess). With the rise of distributed information and knowledge in digital form on the web, the concept of the commons helped “to conceptualize new dilemmas” (Ostrom and Hess 4), given the way new information and communication technologies have affected “how knowledge is managed and governed, including how it is generated, stored, and preserved” (Ostrom and Hess 9). Ostrom and Hess emphasized the fact that these new technologies can promote either “the robustness or vulnerability of a commons”. For them, digital forms made knowledge “more vulnerable than ever before”, enabling the “ability to capture the previously uncapturable” (Ostrom and Hess 10, 14).

When hard-copy journals, for instance, were sold to libraries and individuals, the decentralization of multiple copies made the works robust. When journals are in digital form and licensed to libraries or individuals, the works are centralized and vulnerable to the whims or happenstance of the publisher. (Ostrom and Hess 14)

Therefore, the authors alerted to the importance of keeping the conditions for the preservation and sustainability of knowledge as a common good for present and future generations. They highlighted the emergence of scientific and social movements in favor of knowledge and information commons, as a central pillar of the struggle for democracy. On the other hand, they argued that “knowledge commons is not synonymous with open access”, stressing their understanding that “a commons is a shared resource that is vulnerable to social dilemmas” (Ostrom and Hess 14).

The authors pointed out similarities and differences between knowledge and “traditional commons”. Both were considered goods, resources jointly used and managed by groups at different levels and scales. For them, “the essential questions for any
commons analysis are inevitably about equity, efficiency, and sustainability” (Ostrom and Hess 6). On the other hand, they highlighted the cumulative character, the complex nature and the “dual functionality” of knowledge—“as a human need and an economic good”, “both a social process and a deeply personal process”. They advocated that the IAD framework could be “of value in understanding knowledge as a commons—in regard to both the public-good aspects of this commons and the common-pool resource aspects” (Ostrom and Hess 16).

Understood as global commons, neither a private nor a strict public resource, knowledge was defined as:

... all intelligible ideas, information, and data in whatever form in which it is expressed or obtained... to all types of understanding gained through experience or study, whether indigenous, scientific, scholarly, or otherwise non academic. It also includes creative works, such as music and the visual and theatrical arts. (Ostrom and Hess 7-8)

In this sense, Ostrom and Hess were not only referring to scholarly and scientific knowledge, but to more extended knowledge concepts and issues “far beyond the ivory tower.”

**Alternative Views**

Alternative and complementary perspectives on the common(s) have been developed relative to that proposed by Ostrom and team. We summarize below those that address more directly the aspects that we want to emphasize here.

A first set of arguments refutes a rigid division between natural and intellectual commons, arguing that this distinction should rather be reinterpreted as a matter of emphasis in the analysis, considering the necessary dual character of the common (Vieira). From this perspective, Hardt and Negri claim that their notion of the common:

... does not position humanity separate from nature, as either its exploiter or its custodian, but focuses rather on the practices of interaction, care, and cohabitation in a common world, promoting the beneficial and limiting the detrimental forms of the common. (xviii)

In this sense, Massimo De Angelis recalls the contribution of historian Peter Linebaugh who popularized the term “commoning” as corresponding to “the (re)
production of commons”. David Bollier also refers to Linebaugh when he states that “there is no common good without commoning”, understood as “a set of ongoing practices, not an inert physical resource” (351). He claims that “the commons is not only about shared resources; it’s mostly about the social practices and values that we devise to manage them” (Bollier 351).

A second set of arguments questions the idea of intrinsic characteristics of the commons, considering they are the result of collective and conflicting decisions and actions that promote or hamper practices that ensure their equitable and sustainable management.

For Dardot and Laval:

Nothing is in itself or by nature “common”. Ultimately it is social practices and only them that decide on the “common” character of a thing or a set of things. Therefore, against any naturalism or essentialism it is necessary to maintain that it is the activity of men which makes something a common, keeping it from any logic of appropriation and reserving it for collective use. (271)

A third set of arguments criticizes the idea that more information leads to better politics as concealing the conflicts and inequalities within social relations. On the contrary, social actors usually diverge about expectations and objectives and hardly establish long-lasting agreements based on consensus and mutual truth. The conflicting dimension is an integral part of the commons and its governance. From this perspective, Dardot and Laval argue that:

The conflict dimension must be recognized as part of the common and not considered an unfortunate ‘side effect’ that should be avoided: the common it does not constitute itself, it does not perpetuate itself and it does not expand in any way other than in and through conflict. What is instituted as common is in active opposition to a privatization process (be it urban space, water or seeds). (271)

Finally, Lafuente and Estalella developed a theoretical-conceptual approach in which common does not only mean common goods (the commons), referring to an economic sense. Common also—and mainly—refers to the relationship with otherness, “in between”, in a more anthropological sense. From this perspective, they propose the notion of “common science”, which combines knowledge activism and knowledge production, opening up science agenda, concepts and methods to the scrutiny and contribution of other epistemic groups. This mode of science
acknowledges the epistemic value of the “experiential” as well as the ordinary knowledge. In this sense, lay people should be recognized as “experts in experience” (Callon and Rabeharisoa) who produce relevant knowledge from solving problems in everyday life and from participating in social movements. Actors with different points of view not only alter the social composition of science, but also promote alternative modes of knowledge. Lafuente considers that it favors “more robust” decisions because “each new collective incorporated represents a lower degree of exclusion, implies an extension of freedoms, and, finally, makes visible an expanded society beyond the limits we believed to be insurmountable”6 (144).

A “Situated” View from Ubatuba: Possibilities and Limits of the IAD framework

Our case study on open science in the context of the municipality of Ubatuba helped us to identify both possibilities and limits in the use of the IAD framework in a concrete situation. Among the limits, we point out:

- The difficulty in delimiting a single common pool resource in that region, since our action arena is not restricted to the management of a single resource of common use, but it refers to a medley of resources.

- The larger scale and heterogeneity of perspectives and interests of its population, which hinders trust relations and mutual agreement. While the IAD case studies reported by Ostrom and team encompass small-scale resource systems, usually involving communities with about 50 to 15,000 people who depend heavily on that resources for their livelihoods, Ubatuba municipality has a population of about 80,000 inhabitants, in a heterogeneous composition and diverse interests and conflicts: indigenous communities, fishermen, caïcaras, quilombolas and a multitude of floating residents and seasonal tourists.

- The difficulty to operationalize the IAD framework, since it requires information that is not easily available to the researcher, which would require an intense and prolonged fieldwork, broader than what was feasible for the project.

- Finally, the fact that, as pointed by Vieira, the IAD framework does not sufficiently address considerations about broader sociopolitical and economic relations and long-term historical processes, nor the analysis of power relations, conflicts and inequalities among actors with different expectations. In our case, the economic and

6. “... cada nuevo colectivo incorporado representa un grado menor de exclusión, implica un ensanchamiento de las libertades y, por fin, hace viable una sociedad expandida más allá de los límites que creíamos infranqueables.”
territorial occupation through tourism, real estate, oil and gas industries place local
dynamics in strong interaction and interdependence with broader scales (regional,
national and global).

On the other hand, the IAD framework was useful to think of our research field as
an action situation, in motion, a nonstatic reality, therefore quite consistent with our
perspective of action research.

We defined as our action situation the process of revision of the ecological-economic
zoning (EEZ) of the North Coast of São Paulo, which is a legal responsibility of
the State Government. With the help of the Municipal Department of Environment,
public hearings were held in the region, aiming to inform the population about the
process and mobilize residents to draw up a plan in line with local demands. More
than 80 requests for modifications were submitted by the municipal government,
responding to demands submitted by the local population. The effort of the Tradition-
al Communities Forum (FCT) to develop its own map to support the EEZ revision
reveals the understanding that the role of information users is very limited for those
who aim to interfere politically. These communities wish not only to contribute their
accumulated experience and knowledge in the sustainable management of common
resources but also to influence the decisions that affect them.

Nevertheless, several local leaders reported problems in the traditional communities’
participation in the EEZ revision process. Our research evidenced that overlapping
the roles of policy-makers and information providers is not mere coincidence. It is an
essential strategy for data production that legitimates the arguments competing for
the definitions of the EEZ and, in the end, disputing different conceptions and stra-
tegies of development. In this process, scientific knowledge, produced in research
institutes, universities and laboratories, is mobilized by public policymakers with the
purpose of certifying and, therefore, legitimating their proposals. In this sense, the
mastery of technical language and the ability to translate this knowledge into infor-
mation represented in maps—and, even more so, in rules-in-form—is an important
advantage for policymakers.

We concluded that the role assigned to information and knowledge, from different
points of view and strategies, lies at the heart of disputes between a managerial
logic that also implies the normalization of new forms of control versus a logic of
dissent that disputes the criteria of “measure” of instrumental reason and defines
new instituting forms of the common. And this represents the opposite of the
idea of an “institutional robustness” based on the co-construction of rules-in-use
involving different actors and perspectives. At the end, the perpetuation of the
commons depends on the ability of actors to evolve rules and to build democratic ways to dispute antagonistic views.

These findings are congruent with the criticisms of the open science initiatives focused only on “access”, since they may reinforce the equivalence between a “well informed” policy and a “certified” one. In this sense, we adopted the notion of common science, proposed by Lafuente and Estalella, emphasizing the importance of favoring the interlocution of science with other cognitive actors and their knowledge bases.

Finally, the IAD framework served us as an analytical tool that was instrumental in dealing with different variables and dimensions. Congruent with our territorial perspective, the IAD framework made us critically interrogate the mutual feedback between natural (or more widely, material) and knowledge/information commons. Nevertheless, during our action research project, we aimed to combine, on the one hand the IAD framework’s initial focus on the collective management of natural commons, and on the other hand, its focus on the collective management of knowledge commons. We proposed to reflect on the interaction and co-determination between these two common resources. Moreover, we shed light on the fact that they are not just two different types of commons, but are mainly two dimensions of the same “commoning” process.

Concluding Remarks

From the outset of our project, we questioned the current idea in the open science movement that open access to scientific information would be capable or be sufficient to reduce asymmetries and promote “better informed” and more egalitarian policies. We observed that, although democratization in access to information and knowledge—and even the recognition of the contribution of different modes of knowledge—is important, this does not solve the asymmetries of power over common pool resources. The disputes over the natural commons and the information and knowledge commons are part of a local institutional dynamics characterized by:

- the flow of public, private and nongovernmental interests through different spaces, within an informal network of influence;
- the existence of a complex movement of alliances and tensions between actors and their different institutional positions;
• the diversity of forms of information and knowledge production and demands as a central aspect of the positioning of the actors in affirming and defending their views and claims with respect to the commons.

In this sense, we adopted a substantive socio-territorial approach, relying on the Brazilian geographer Milton Santos’s conception of space as a hybrid made up of the indissociable union of systems of objects and systems of actions (Santos; Albagli). The sustainable and equitable management of the natural local resources as a commons is inextricably co-related to the disputes over the production and circulation of knowledge and information as a commons. The production of knowledge commons involves the social appropriation of the territory, as well as a common material base, composed of natural and artificial goods that support life in common. On the other hand, the so-called “exogenous” variables in the IAD framework—biophysical characteristics, attributes of the community, and rules-in-use—are an intrinsic part of an action situation.

Our common science approach, based on Lafuente and Estalella, implied the recognition that a pluralistic ecosystem of knowledge modes is closely related to a pluralistic ecosystem of modes of existence (Albagli et al.). From this perspective, in our case, we considered the territory as the very infrastructure of the production and reproduction of knowledge common, both as the material basis of life in common and as the space where modes of subjectivation for the commoning takes place.

References


# Table of contents

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td><em>Leslie Chan</em></td>
<td></td>
</tr>
<tr>
<td>Vertical Integration in Academic Publishing</td>
<td>15</td>
</tr>
<tr>
<td><em>George Chen, Alejandro Posada, Leslie Chan</em></td>
<td></td>
</tr>
<tr>
<td>The End of a Centralized Open Access Project and the Beginning of a Community-Based Sustainable Infrastructure for Latin America</td>
<td>41</td>
</tr>
<tr>
<td><em>Arianna Becerril-García, Eduardo Aguado-López</em></td>
<td></td>
</tr>
<tr>
<td>Global Scholarly Collaboration</td>
<td>57</td>
</tr>
<tr>
<td><em>Sergey Parinov, Victoria Antonova</em></td>
<td></td>
</tr>
<tr>
<td>The Value of Network Sustainability</td>
<td>77</td>
</tr>
<tr>
<td><em>Elisabeth Ernst</em></td>
<td></td>
</tr>
<tr>
<td>Whose Infrastructure? Towards Inclusive and Collaborative Knowledge Infrastructures in Open Science</td>
<td>97</td>
</tr>
<tr>
<td><em>Angela Okune, Rebecca Hillyer, Leslie Chan, Denisse Albornoz, Alejandro Posada</em></td>
<td></td>
</tr>
<tr>
<td>Availability of Cultural Heritage Structured Metadata in the World Wide Web</td>
<td>121</td>
</tr>
<tr>
<td><em>Nuno Freire, Pável Calado, Bruno Martins</em></td>
<td></td>
</tr>
<tr>
<td>An Expertise Recommender System based on Data from an Institutional Repository (DiVA)</td>
<td>135</td>
</tr>
<tr>
<td><em>Milena Angelova, Vishnu Manasa Devagiri, Veselka Boeva, Peter Linde, Niklas Lavesson</em></td>
<td></td>
</tr>
<tr>
<td>The Public Knowledge Project</td>
<td>151</td>
</tr>
<tr>
<td><em>Juan Pablo Alperin, John Willinsky, Brian Owen, James MacGregor Alec Smecher, Kevin Stranack</em></td>
<td></td>
</tr>
<tr>
<td>Beyond the Dichotomy between Natural and Knowledge Commons</td>
<td>165</td>
</tr>
<tr>
<td><em>Sarita Albagli, Anne Clinio, Henrique Z.M. Parra, Felipe Fonseca</em></td>
<td></td>
</tr>
</tbody>
</table>
Réalisé sur les presses de l’imprimerie Ciaco
en impression à la demande
depuis le mois de mai 2019
à Louvain-la-Neuve, Belgique
N° d’imprimeur : 89470