Reassembling Scholarly Communications

Histories, Infrastructures, and Global Politics of Open Access

OA Funding Provided By:

- Arcadia Fund
- Birkbeck, University of London

The open access edition of this book was made possible by generous funding from Arcadia—a charitable fund of Lisbet Rausing and Peter Baldwin.

The title-level DOI for this work is:

doi:10.7551/mitpress/11885.001.0001

20 Toward Linked Open Data for Latin America

Arianna Becerril-García and Eduardo Aguado-López

Scholarly communication is perhaps *the* phase in the research life cycle that has most seized the opportunity to broaden inclusion through the use of information technologies. Open access has promoted free and unrestricted access to scientific content, especially, driven by mandates, when it has been publicly funded. OA holds out the promise of a global scientific dialogue that would allow for a more inclusive, global research ecosystem.

Globalization has indeed become the ultimate goal in scientific practice, in which the circulation of knowledge generated in all regions is expected to have worldwide visibility. Often, this goal of global visibility has been equated with journals' presences in "mainstream" databases such as Web of Science (WoS) or Scopus. Those outside the Global North are encouraged to publish in journals indexed by these databases if their contributions are to have international visibility (although this is not guaranteed), but also so that these publications are viewed as high quality.[1]

Latin America, as with many other developing regions, has historically faced a lack of visibility and recognition for the science that it generates. This is mainly due to the scarce presence of Latin American journals in the aforementioned mainstream databases, which has led to the marginalization of research produced in the region.

Indeed, only 276 Latin American journals are indexed by WoS and 795 by Scopus, whereas in Redalyc there are 1,111. Figure 20.1 shows a Venn diagram with the journal sets' distribution among Redalyc, WoS, and Scopus. Further, a deeper analysis shows that most of the few indexed journals hold very low quartile positions. This distorted representation is not spread evenly between the disciplines. For instance, the social sciences and humanities (SSH) are particularly poorly represented. Only 90 social science

and humanities journals from this region are indexed by WoS and 361 by Scopus. However, Redalyc indexes 555 journals from those areas (see figures 20.1 and 20.2).

This paradigm of valuation and communication presents a conundrum for the regional context. That is: there is low representation of Latin American research output in the legitimated knowledge circulation channels for the Global North, even though this region is possessed of an extremely robust ecosystem of science communication—and a system that is natively open and scholar-owned at that. Indeed, Latin American scholarly journals are led, owned, and financed by academic institutions. As covered in other chapters in this volume, each academic institution is part of an informal cooperative system that is neither formalized nor made explicit. Each institution supports journals that are managed by their own faculty members and the content of these journals is available to everyone. Where an institution is publicly funded, public budgets from local or national governments are used to support these publications. In this way, each institution's investment in journals mutually benefits all other institutions. This kind of

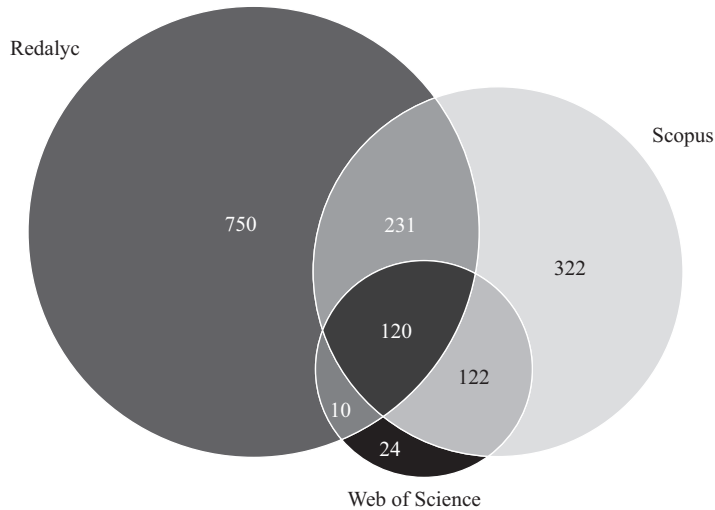

Figure 20.1
Latin American journals indexed by Redalyc, Scopus, and WoS.
Data sources: Redalyc database (2018), Scopus Source Title (2018), Source Publication List for Web of Science: Science Citation Index Expanded (2017), Social Sciences Citation Index (2017), Arts & Humanities Citation Index (2017).

informal cooperative was already operational before the term "open access" was even coined.

This Latin American ecosystem is composed of several layers. The base level is supported by hundreds of "university presses" with journals published electronically using software such as Open Journal Systems. Then, in an upper layer, platforms such as CLACSO, Redalyc, SciELO, and Latindex provide a set of added value features. Latindex's job, for instance, is to keep a well-organized directory of quality journals published in the region. CLACSO has contributed strongly to the Open Access Movement with promotion of and contents for the social sciences. Redalyc provides journals with mechanisms to increase their visibility, services of interoperability, search engine optimization, metrics, usage tracking, and more recently, technology to procure XML typesetting under the JATS (Journal Article Tag Suite) standard, then transformed automatically to PDF, HTML, and EPUB file formats of articles.[2]

Latin America has relied upon open access as its path to inclusion in a more participatory worldwide scholarly system. Originally, with the OA initiatives and declarations, a counterweight was sought to reduce the

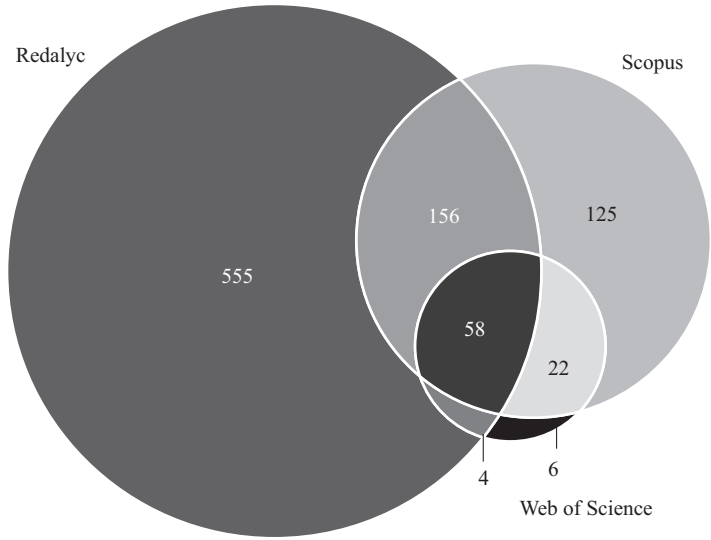

Figure 20.2
Latin American Social Sciences and Humanities Journals Indexed by Redalyc, Scopus, and WoS.

asymmetries generated by the primary communication, collaboration, and dissemination channels in the Global North. As noted by Marin, Petralia and Stubrin, and Banerjee, Babini, and Aguado, OA is viewed as the best option to promote a democratic and inclusive development and has proven results in increasing the international visibility of research.[3]

Yet, this has been shown to be an overly optimistic stance. For although, as highlighted by Babini, open access is the standard in Latin America, this openness has not broken the inertial dependencies of traditional legitimation circuits.[4] Thus, the exclusion, asymmetry, and gaps remain.

Further, this regional OA landscape is threatened by commercial open-access strategies from the Global North, which put at risk of rupture the Latin American OA nonprofit ecosystem while proposing to move to a new circumstance of exclusion: from "paying to read" to "paying to publish" (the APC-based OA model).

Hence, openness is not enough. It remains imperative also to modify systems of research assessment and to find more effective methods of communicating the knowledge generated in different regions, disciplinary fields, and languages. As Beigel suggests, it is not about giving the voices from the South a space in the channels where the North is established, but to question the very foundations of supposedly "universal" academic recognition and find ways to implement a non-hegemonic transnational dialogue.[5]

There are multiple approaches to achieving this. One strategy in Latin America is gambling upon reaching visibility within existing legitimized channels by adopting questionable research assessment practices, such as the use of the impact factor. This is the approach adopted by the SciELO Citation Index. Conversely, others such as Redalyc and CLACSO seek to integrate the region's developments, experience, and the academic model in order to minimize costs and join forces to guarantee the sustainability of OA and to maintain the academic-owned nature of dissemination and production of knowledge. This is being done through a recently launched, initiative called AmeliCA (Open Knowledge for Latin America and the Global South), which is supported by UNESCO and dozens of universities throughout the region.[6]

Technology for Visibility, Discoverability, and Internationalization

Some of the questions that arise when trying to build a more neutral, equitable, and inclusive space for scholarly communications include: are

technologies capable of contributing to this? What might be the roles of semantic technologies, artificial intelligence techniques, ontological engineering, natural language processing, machine learning, and other advancements? We believe that there is a future role for technological innovations to contribute to a more integrated knowledge ecosystem and here go on to describe the semantic technologies that could help, without adopting a wholesale techno-solutionist perspective.

Certainly, interoperability is an important area in which technological developments have already been applied. The concept of interoperability arose from the need to exchange information across different applications and organizations with diverse data sources. What, though, if interoperability principles could be applied to scholarly communication in terms of the interchange of research results across geographical regions, disciplines, or even languages? Research published online—particularly when it is openly accessible—has the potential to join a giant mass of knowledge where visibility and discoverability are achieved intrinsically. A researcher from any place could retrieve any informational input needed to do his or her job and, eventually, his or her results would rejoin this database. Everything starts, though, with data structuring.

On the web, scholarly resources have been structured by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for interoperability purposes. In turn, this has contributed to the visibility of contents because metadata can be automatically distributed to libraries, universities, portals, and aggregators in ways that facilitate retrieval and consumption.

The data model specified by OAI-PMH provides a basic semantic level for understanding the nature of described resources, but only at an identification level. This is insufficient fully to capitalize on all textual elements, including citation data, figures, mathematical expressions, tables, supplementary material, and more.

Having scholarly resources structured at the element level goes well beyond OAI-PMH capabilities. This is an area where eXtensible Markup Language (XML) plays a major role, since it provides a set of simple rules and a uniform method to describe and exchange structured data, separated from the format in which the information is presented. XML—of which JATS is a schema—enables the structuring of full texts of scholarly resources and brings them a greater potential for readability and indexing, which favors their capacity to be discovered. It also, as Martin Paul Eve outlines

in his chapter, facilitates potential future machine-reading possibilities for ingesting the scholarly corpus.

As Abel Packer points out elsewhere in this book, SciELO has promoted the use of XML since 2012 but began its full-scale adoption across all of its journals as of 2015. Health sciences journals began to adopt it as of 2014.[7] Meanwhile, Redalyc started to adopt XML in 2015 with a strategy based on the empowerment of scholarly publishers, providing tools and knowledge to make XML tagging a sustainable process.[8] Currently, approximately 90 percent of journals indexed by Redalyc publish their content in XML JATS.

While the implementation of XML in journals carries great potential, there is a deeper and more relational level of granularity at which information could be disseminated. Every piece of information that comprises a text from a journal article or from any other scholarly content could be understood, interpreted, and linked into a "knowledge cloud."

There are many barriers to such a global system, though. As noted by Ora Lassila, although everything on the web is machine-readable, it is not machine-comprehensible.[9] For instance, the information content of scholarly outputs could be represented as connections of informational elements where the structure, formed by nodes and connections, expresses knowledge. That form of structuration, though, goes far beyond the capabilities of XML, whose data model is a tree. Indeed, we would argue that a far better data model for knowledge representation is a graph, as provided by RDF (a resource description framework).

Thus, we argue, a transition needs to be made from a machine-readable to machine-comprehensible paradigm with respect to scholarly information resources: a transition from XML to RDF.

Leveraging Semantic Technologies to Achieve a Global Research Dialogue

The "HowOpenIsIt?®" Open Access Spectrum guide provides a scale for machine readability of OA content that includes, as a maximum level of openness, a notion of semantics that has not yet been achieved by Latin American journals.[10] RDF, the technology that would enable this, is an abstract model, a way to break down knowledge into discrete pieces.[11] And, indeed, there are two different purposes behind XML and RDF that should

be understood for a future semantic scholarly context. This boils down to the use cases: for those who wish to query documents (XML) and those who wish to extract the "meaning" in some form and query that (RDF).[12]

Minimal structuring and semantics are integral to the web as it currently exists, in the form of hypertext. The essential feature of hypertext is the nonlinearity of content production by the authors and of content perception and navigation by users.[13] Indeed, from even minimal semantics have arisen amazing results. What, though, if web pages had more semantics?[14] Semantics, the process of communicating enough meaning to result in an action, has great potential to enable scholarly resources to join the so-called Web of Data.[15]

Semantic technologies discover relationships that exist among resources and then represent those relationships via some form of metadata, making it easier to develop reusable techniques for querying, exploring, and using the underlying data.[16] Using this semantic web, software can process content, reason with it, combine it, and perform deductions logically to solve problems automatically.

We, the authors of this chapter, have previously applied semantic technologies to structured scholarly resources. The results consist of a semantic model for selective knowledge discovery dubbed "OntoOAI" a semantic application that enables the processing of data structured with OAI-PMH, the application of ontologies in the description and verification of the knowledge obtained from OAI-PMH resources, and inference-testing mechanisms on the resultant dataset.[17]

OntoOAI was executed using a combination of three sources of information: Redalyc, the institutional repository of Roskilde University (RUDAR), and DBpedia. This data integration was possible through two ontologies: Dublin Core and Friend of a Friend (FOAF). OntoOAI processed 395,940 items resulting in 7.9 million triplets, which correspond to granular pieces (for instance, 60,354 triplets of author names; 1.6 million triplets of topics; 394,775 triplets of dates, and more).

It should be noted that given the identified associations between resources, it is possible to take advantage of graphs, hierarchical, or other net visualizations that allow users to explore and browse information following relations at different levels, which adds value for discoverability purposes.

OntoOAI's application verified the feasibility and benefits of using semantic technologies to achieve selective knowledge discovery while also

showing some of the limitations of using OAI-PMH data for this purpose (among which is the lack of both URIs and full-text structuration). The latter would enable a journal article (or another scholarly resource) to be broken down into pieces that individually would form nodes in a graph whose relations among them are represented as edges and together they might be expressed in an ontology. RDF based on JATS could also work to achieve that task (see figure 20.3). Indeed, if this lack of URIs and RDF availability are overcome by Latin American scholarly resources, all this information could be part of the Linked Open Data (LOD) Cloud.[18] This would mean that every piece of information published by scholarly journals in Latin

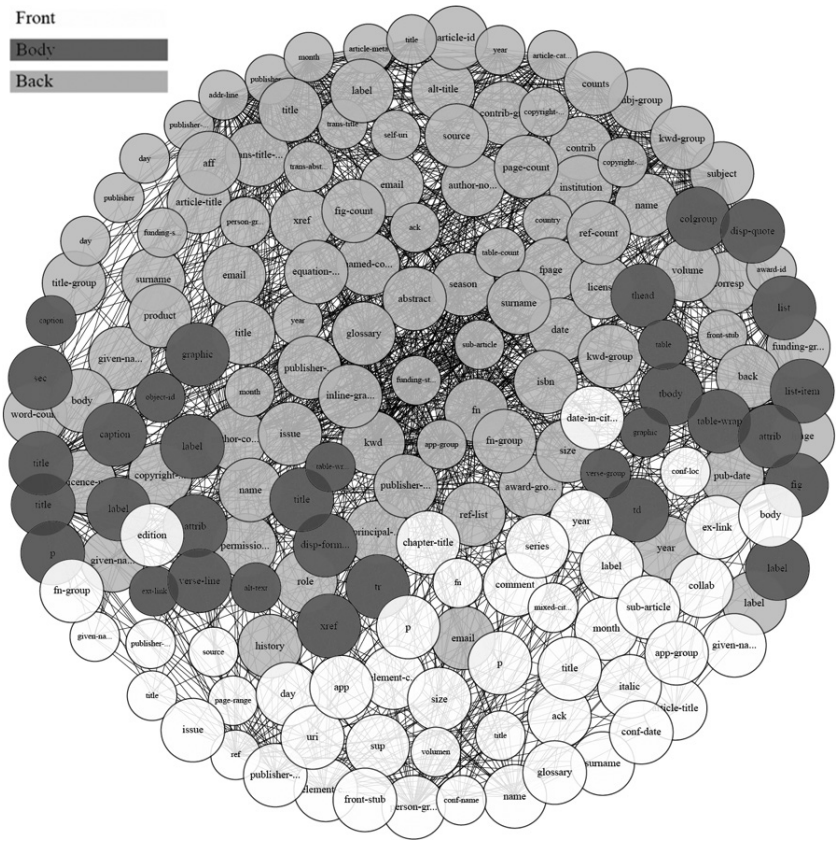

Figure 20.3
Knowledge representation of a journal article (RDF derived from JATS XML) based on the representation of the Linked Open Data Cloud.

America could be linked to all data provided by all other LOD sources (see figure 20.4). Had we such semantic markups within our systems of scholarly communications, novel mechanisms of knowledge discovery could be developed to query, extract, infer, and retrieve information in such a way that usability and applicability of knowledge generated in Latin America—and other regions—could be improved, and that published knowledge *per se* could reach visibility, discoverability, and internationalization, all provided by the inherent composition of it in the knowledge structure. Thus, traditional circuits of scholarly communication, the ones legitimated by current research assessment strategies, could be left behind. Information could speak by itself in benefit of a global science communication.

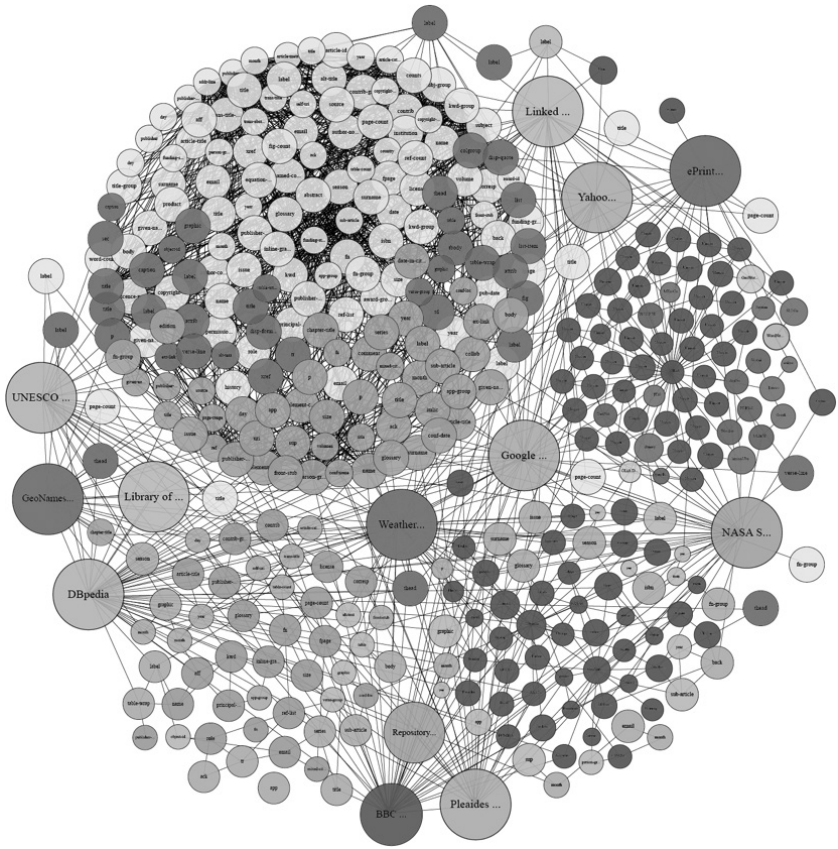

Figure 20.4
Journal articles as part of the Linked Open Data Cloud.

Certainly, many will see this technological solution as overly optimistic. After all, most difficult problems have social, rather than technological, answers. Yet we believe in the potentially liberatory powers of information technologies.

Notes

1. Eduardo Aguado-López, Arianna Becerril-García, and Sheila Godínez-Larios, "Colaboración Internacional en Las Ciencias Sociales y Humanidades: Inclusión, Participación e Integración," *Convergencia Revista de Ciencias Sociales*, no. 75 (2017): 16, https://doi.org/10.29101/crcs.v0i75.4227.

2. JATS, the international standard ANSI/NISO Z39.96–2015, defines a set of XML elements and attributes that describe content and metadata of journal articles, aimed to provide a common format in which journal content can be exchanged. National Information Standards Organization, "JATS: Journal Article Tag Suite, Version 1.1," 2015.

3. Anabel Marin, Sergio Petralia, and Lilia Stubrin, "Evaluating the Impact of Open Access Initiatives within the Academia and Beyond," in *Made in Latin America: Open Access, Scholarly Journals, and Regional Innovations*, ed. Juan Pablo Alperin and Gustavo Fischman (Ciudad Autónoma de Buenos Aires: CLACSO, 2015), 75–102, http://biblioteca.clacso.edu.ar/clacso/se/20150921045253/MadeInLatinAmerica.pdf; Dominique Babini, Eduardo Aguado López, and Indrajit Banerjee, "Tesis a Favor de La Consolidación Del Acceso Abierto Como Una Alternativa de Democratización de La Ciencia En América Latina," in *Acceso Abierto*, by Peter Suber (México: Universidad Autónoma del Estado de México, 2015), 13–48.

4. Dominique Babini, "Voices from the Global South on Open Access in the Social Sciences," in *Open Access Perspectives in the Humanities and Social Sciences* (London: London School of Economics, 2013), 15, https://blogs.lse.ac.uk/impactofsocialsciences/files/2013/10/Open-Access-HSS-eCollection.pdf.

5. Fernanda Beigel, "El Nuevo Carácter de la Dependencia Intelectual," *Cuestiones de Sociología* 14 (2016): 9, http://hdl.handle.net/10915/54650.

6. Redalyc, CLACSO, and UNESCO, "AmeliCA—Conocimiento abierto para América Latina y el sur Global," 2019, http://www.amelica.org/; see also "AmeliCA vs Plan S: Same Target, Two Different Strategies to Achieve Open Access.—AmeliCA," accessed May 1, 2019, http://www.amelica.org/en/index.php/2019/01/10/amelica-vs-plan-s-mismo-objetivo-dos-estrategias-distintas-para-lograr-el-acceso-abierto/.

7. SciELO, "¿Porqué XML?," *SciELO En Perspectiva* (blog), April 4, 2014, https://blog.scielo.org/es/2014/04/04/porque-xml/.

8. Eduardo Aguado-López, Arianna Becerril-García, and Salvador Chávez-Ávila, "Conectando al Sur Con La Ciencia Global: El Nuevo Modelo de Publicación en

ALyC, No Comercial, Colaborativo y Sustentable," 2016, 8–10, https://blogredalyc.files.wordpress.com/2016/08/redalycnuevomodelopublicacion2016-11.pdf.

9. Ora Lassila, "Web Metadata: A Matter of Semantics," *IEEE Internet Computing* 2, no. 4 (1998): 1, https://doi.org/10.1109/4236.707688.

10. Scholarly Publishing, Academic Resources Coalition, Public Library of Science, and Open Access Scholarly Publishers Association, "HowOpenisit?," Public Library of Science, 2014, https://www.plos.org/files/HowOpenIsIt_English.pdf.

11. Joshua Tauberer, *What Is RDF and What Is It Good For?* (2014; repr., Github, 2008), https://github.com/JoshData/rdfabout.

12. Tim Berners-Lee, "Why RDF Is More Than XML," W3C, September 1998, https://www.w3.org/DesignIssues/RDF-XML.html.

13. Gerti Kappel et al., "An Introduction to Web Engineering," in *Web Engineering: The Discipline of Systematic Development of Web Applications*, ed. Gerti Kappel (Hoboken, NJ: John Wiley & Sons, 2003), 11.

14. Steve Bratt, "Semantic Web, and Other Technologies to Watch," W3C, January 2007, https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(1).

15. Toby Segaran, Colin Evans, and Jamie Taylor, *Programming the Semantic Web* (Sebastopol, CA: O'Reilly Media, 2009), 3.

16. Oswald Campesato and Kevin Nilson, *Web 2.0 Fundamentals: With AJAX, Development Tools, and Mobile Platforms* (Sudbury, MA: Jones & Bartlett Learning, 2010), 33; Segaran, Evans, and Taylor, *Programming the Semantic Web*, 37.

17. Arianna Becerril-García and Eduardo Aguado-López, "A Semantic Model for Selective Knowledge Discovery over OAI-PMH Structured Resources," *Information* 9, no. 6 (2018): 4–12, https://doi.org/10.3390/info9060144; Arianna Becerril-García, Rafael Lozano Espinosa, and José Martín Molina Espinosa, "Semantic Approach to Context-Aware Resource Discovery over Scholarly Content Structured with OAI-PMH," *Computación y Sistemas* 20, no. 1 (2016): 131–135, https://doi.org/10.13053/cys-20-1-2189; Arianna Becerril-García, Rafael Lozano Espinosa, and José Martín Molina Espinosa, "Modelo Para Consultas Semánticas Sensibles al Contexto Sobre Recursos Educativos Estructurados con OAI-PMH" (Encuentro Nacional de Ciencias de la Computación, ENC 2014, Oaxaca, Mexico, 2014), 1–4.

18. The Linked Open Data Cloud is available at https://lod-cloud.net.