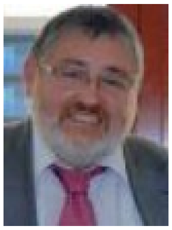
Low visibility of Latin American repositories in Google Scholar: technical incompatibility or lack of web strategy?

blogs.lse.ac.uk/impactofsocialsciences/2014/07/31/latin-american-repositories-google-scholar-low-visibility/

7/31/2014

The content in many repositories in Latin America fail to come up in systematic searches largely due to the inadequate use of domain names and metadata schema, find **Enrique Orduña-Malea** *and* **Emilio Delgado-López-Cózar**. *Institutional repositories are ultimately websites and concepts such as usability, information architecture, search engine optimization, among others, should be considered in their primary design. In a context like Latin America, in which scholarly production requires extra visibility because it lies outside the academic mainstream, repositories are essential vehicles, and their low visibility could significantly affect their real use by end users.*



Informal communication among scientists has been produced since ancient times, long before the Open Access term was coined. The official launch of institutional repositories (IR) simply involves an institutionalisation of the academic material exchanging practice within an institution. Apart from facilitating the primary functions of providing deposit and access to the entire body of knowledge of a university, the IR additionally assists scholars in managing their own digital resources, thus constituting a critical component of the academic communication system.

However, after the massive implementation of IRs in universities around the world, if holdings and services provided at present are considered, beyond the mere marketing of academic research, their true purpose remains unclear.

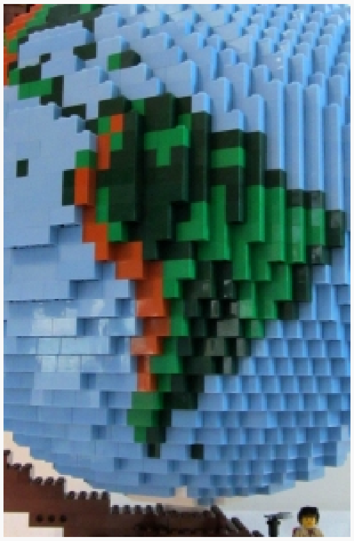


Image credit: dirkb86 (Flickr, CC BY 2.0)

Indeed, some difficult questions remain to be answered, such as:

- When can we say an IR is fulfilling its function satisfactorily?
- How can we prove that IRs are valuable vehicles to disseminate scientific knowledge themselves?

If we exclude the value of the items (usually published on journals outside the repository), which is a value that belongs to the item itself, and does not to the repository itself, this becomes rather complex.

The main problem that prevents us answering the above questions easily is that the creation and dissemination of repositories generally takes the "Open Access" concept as its centre of gravity, forgetting that the product is, at the end of the day, a website. And concepts such as usability, information architecture, search engine optimization, among others, should be considered in its primary design, as it happens in the creation of any other web page.

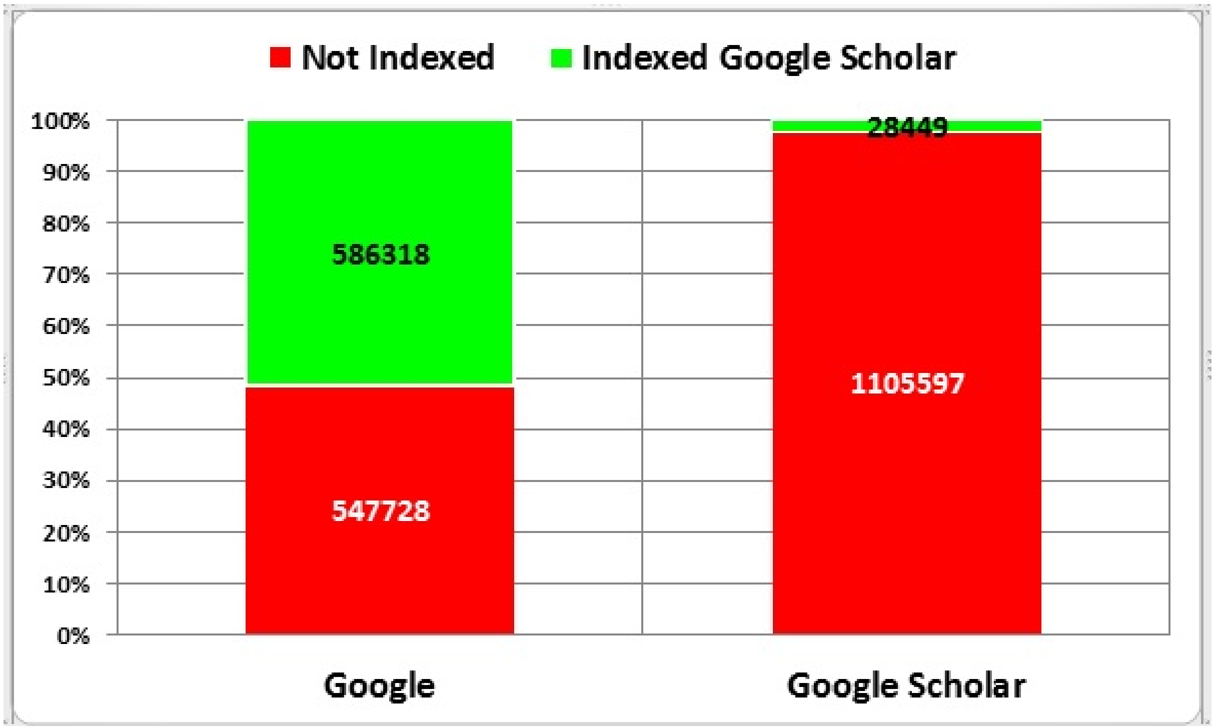
Let us consider a simple but illustrative example: according to *OpenDOAR* – the Directory of Open Access Repositories (as of July 2014), there are over 291 repositories in Latin American countries (considering South America, Central America, The Caribbean, and Mexico). Notwithstanding, in the last edition of the Ranking Web of Repositories (July 2014), we find only 170 repositories. Why?

This is because of the mainly inadequate use of domain names. The creation of subdirectories is legitimate, and it works on a technical level, but this jeopardised the task of measuring accurately. It is therefore a wrong strategic decision. Otherwise, the indexation of repositories in the search engines (especially in Google and Google Scholar) is even more important. A repository whose contents are not indexed in the most important search engine in the world becomes a contradiction.

Bearing in mind the importance of self-archiving in the generation of OA content, the web presence and visibility of the repositories on the Web are essential to ensure that the content they host makes OA truly effective for the community. Arlitsch and O'Brian, in a pioneering work carried out in 2012, detected unexpectedly low indexing ratios (30%) for 21 repositories in the United States. Without going into the possible causes of this phenomenon yet, the effects are devastating: the practical invisibility of contents hosted.

Taking this as a starting point, we considered the following question: what will happen if this phenomenon also occurs in places where repositories are key vehicles to spread outside their countries the science generated outside the mainstream? Let's be clear, is it the same if a North American repository is invisible to Google as it is a South American repository?

With this purpose, we figured out an empirical way to ascertain the web presence and visibility of Latin American repositories in Google and Google Scholar through the application of web indicators and operators (mainly the "site" command). For a sample of 127 repositories from 13 countries, the results confirmed that the indexing ratio was low in Google, and virtually nonexistent in Google Scholar.



Percentage of documents (PDF files) from 127 Latin-American Repositories indexed in Google and Google Scholar). Data source: re-elaborated from Orduña-Malea & Delgado López-Cózar (in press)

In any case, the overall data should be viewed with some caution:

- First, because the "site" operator does not return all the items that Google Scholar indexes for a repository. It is neither exhaustive nor accurate.
- Second, because the procedure of grouping multiple versions of an article operates in such a way that one version is taken as the "primary" version. The effect of this procedure in the "site" operator is still to be determined.

Arlitsch and O'Brian also found another main cause: the metadata schema used. This forces us finally, to read the instruction manualof Google Scholar, where the requirements and recommendations for repository webmasters are included.

Among others, we find the following:

- "Each file must not exceed 5MB in size". This is the reason why Google achieved more optimal results than Google Scholar. We already know where all doctoral theses ended up…
- "Use Dublin Core tags as a last resort". Therefore, if Google Scholar, first academic search engine in the world today (in crescendo) clearly tells us "do not use Dublin Core", why the vast majority of metadata repositories offer only Dublin Core?

In short, we do not read the instructions, and worse, we did not even realize about our invisibility. As recently pointed out Pablo de Castro in the INCYT forum, the fault is neither Google nor Dublin Core, but:

a) The difficulty of ensuring that the guidelines are assumed by all repository managers.

b) The discussions about open access traditionally step in ideology instead of exchanges on technical aspects, so that same arguments are repeated decade after decade everlasting way.

Namely, problems are purely technical although much of their solutions require strategic decisions, and should be addressed in the short term to ensure the visibility of repositories, to which institutions are now devoting significant financial and human resources. Although this phenomenon is happening both in the North and the South, the effects of invisibility are different depending on where they occur.

In a context like Latin America, in which scholarly production requires extra visibility because it lies outside the academic mainstream (i.e. published in journals neither indexed in WoS nor Scopus), repositories are essential vehicles, and their low visibility could significantly affect their real use by end users, given the importance of Google (and Google Scholar) to the search and use of academic information today.

Featured Image credit: Lattre, Jean, *Atlas Moderne ou Collection de Cartes sur Toutes les Parties du Globe Terrestre*, c. 1775. (Wikimedia, Public Domain)

Note: This article gives the views of the authors, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Authors

Enrique Orduña-Malea has a MA and PhD in documentation, Master's degree in multichannel contents management, and technical telecommunications engineer from the Polytechnic University of Valencia (UPV). He is specialized in Web metrics, particularly in the creation, diffusion and consumption of contents and products on the Web. He currently works as postdoctoral researcher and teacher at the UPV, and pertains to the EC3 Research Group.

Emilio Delgado-López-Cózar is Professor of Research Methodos at the University of Granada (Spain) and founder of the EC3 Research Group (Science and scientific communication evaluation) specialized on the field of bibliometrics and research evaluation. He is behind the creation of numerous tools for scientific evaluation in the Spanish environment (such as IN-RECS, IN-RECJ, IN-RECH, I-UGR Ranking of Spanish Universities or H Index Scholar, CIRC (Integrated Classification of Scientific Journals), EC3 Metaranking Spanish Universities among others).