

Are Latin-American repositories invisible on Google and Google Scholar?

Enrique Orduña-Malea¹, Alberto Martín-Martín², Juan Manuel Ayllón²,
Emilio Delgado López-Cózar²

¹ EC3: Evaluación de la Ciencia y de la Comunicación Científica, Universidad Politécnica de Valencia (Spain)

² EC3: Evaluación de la Ciencia y de la Comunicación Científica, Universidad de Granada (Spain)

A DIGEST OF

Orduña-Malea, E., & Delgado Lopez-Cozar, E. (2014). *The dark side of Open Access in Google and Google Scholar: the case of Latin-American repositories*. *Scientometrics* (in press)

Preprint available: <http://arxiv.org/abs/1406.4331>



Complementary material available: <http://hdl.handle.net/10481/32271>

SUMMARY

In this issue, not without some embarrassment, we digest a contribution from our own. The main objective of this study is to ascertain the presence and visibility of Latin American repositories in Google and Google Scholar through the application of page count and visibility indicators. For a sample of 127 repositories, the results indicate that the indexing ratio is low in Google, and virtually non-existent in Google Scholar. A complete lack of correspondence between the repository records and the data produced by these two search tools are indicated as well. These results are mainly attributable to limitations arising from the use of description schemas that are incompatible with Google Scholar (repository design) and the reliability of web indicators (search engines). We conclude that neither Google nor Google Scholar accurately represent the actual size of open access content published by Latin American repositories; this may indicate a non-indexed, hidden side to OA, which could be limiting the dissemination and consumption of open access scholarly literature.

KEYWORDS

Google Scholar / Google / Open Access / Repositories / Web indicators / Web visibility / Indexing / Webometrics / Latin America

 <p>Grupo de Investigación EC3 Evaluación de la Ciencia y de la Comunicación Científica</p>	 <p>GOOGLE SCHOLAR DIGEST Research on Google Scholar Empirical evidences <i>Contra data non arguenda</i></p>	<p>EC3's Document Serie: EC3 Google Scholar's Digest Reviews N° 3 Document History Version 1.0, Published on 20 June 2014, Granada</p>
<p>Cited as Orduña-Malea, E.; Martín-Martín, A.; Ayllón, J.M.; Delgado López-Cózar, E. (2014). <i>Are Latin-American repositories invisible on Google and Google Scholar?</i>. EC3 Google Scholar Digest Reviews, n. 3.</p>		
<p>Corresponding author Emilio Delgado López-Cózar: edelgado@ugr.es, Enrique Orduña-Malea: enorma@upv.es</p>		

1. DIGEST

RESEARCH QUESTIONS

- How many Latin American institutional repositories documents are indexed by Google and Google Scholar?
- What is the web impact of content published in Latin American institutional repositories?
- Is there correlation between page count and web visibility?

METHODOLOGY

Unit analysis

Latin American institutional repositories listed in the Ranking Web of Repositories (July 2013 edition)

Sample

127 Latin American institutional repositories

Design

- The size of the repositories in number of items hosted is obtained from the information provided by the platform itself.
- The total number of items listed in Google and Google Scholar is calculated with the command site (<site:domain.com>; <site:domain.com filetype:pdf>).
- Mention values were obtained from the search engine Open Site Explorer. This retrieves both the number of external links for each repository (measured at the aggregate domain level, i.e., all external links from the same domain are counted only once), and the MzRank indicator at subdomain level, which provides an estimated value for the popularity of the websites analysed.
- Additionally, the number of mentions for each URL was calculated from Google, which gave an estimated indicator of the number of external links (<"domain.com" –site:domain.com –inurl:domain.com>)
- A correlation analysis was conducted for all indicators (given the unequal distribution of web data, the Spearman correlation coefficient was applied) as was a principal component analysis (PCA).

Measures

- Number of documents hosted by the repository (ITE)
- Number of files indexed in Google (Gtot)
- Number of PDF files indexed in Google (Gpdf)
- Number of files indexed in Google Scholar (GStot)
- Number of PDF files indexed in Google Scholar (GSpdf)
- Number of times the URL is mentioned (URL)
- Number of external links grouped by domain (V)
- Link popularity score (0 to 10) (Mz)

Period analyzed: All

Data collection date: October 2013

RESULTS

Errors in the functionality of search engines

Page count values for the repository are lower than those shown for the search engines (these errors vary according to the source).

- In the case of Google, 109 URLs whose size is greater than the number of items were located. For PDF values, the number of URLs with this error is lower at 47, which indicates that this query is more accurate than that for overall size. It therefore seems clear that the search engine is retrieving not only items from the repository but also other files hosted on the domain (including those pertaining to the application used to manage the repository).
- In the case of Google Scholar, there are even fewer errors. Total page count yields 11 URLs with page count values greater than those for the repositories, while for PDF files there are only three). In this case, the errors are directly related to errors in the indexing of resources, but they are practically non-existent and are, in any case, detectable and easily controlled.

Google and Google Scholar Coverage

- If we circumscribe the coverage analysis to only those documents in PDF files, a low coverage in Google (48.3%) and virtually nonexistent in Google Scholar (2.5%) is detected (Figure 1).

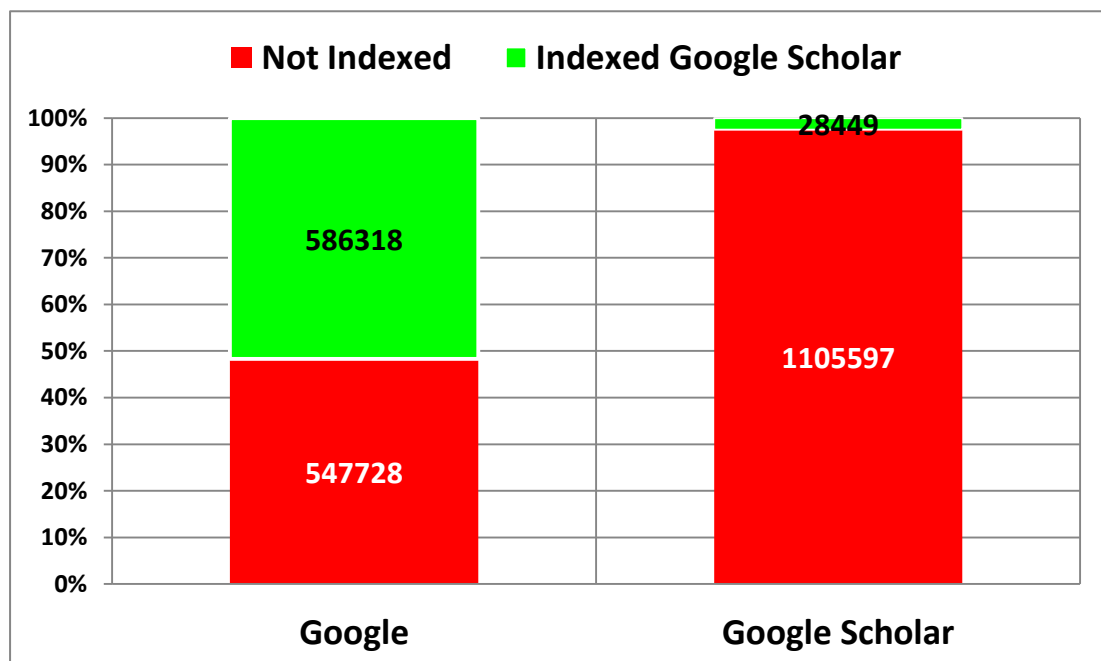


Figure 1. Percentage of documents (PDF files) from 127 Latin-American Repositories indexed in Google and Google Scholar)

Data source: re-elaborated from Orduña-Malea & Delgado López-Cózar (in press)

- If the search is extended to all file types indexed on Google and Google Scholar, the results indicate that Google indexes 100% of the documents and Google Scholar only 34.2%.

Impact of the repositories: mention indicators

- URL mentions: the values obtained are exceptionally high, especially for <tesis.usp.br> (5,380,000 hits). Although search engines round up these values, it is evident that extra noise is high, despite using the <-inurl> command to exclude certain types of spam. Even so, we detected some exceptions in some URLs, which, despite having high page count values (for items both in the repository and indexed by Google), made hardly any impact in URL mentions.
- Referring domains: the achieved impact was very low: only 4 URLs achieved more than 100 domains linking in, while 21 did not return any result. These data correspond to the MzRank values (which depend directly on the quantity and quality of inbound external links on the analysed websites). In this case, no URL scored more than 5 points (the maximum is 10). Moreover, 23 URLs obtained a “0” value.

Correlation between page count and impact

- The number of items retrieved directly from the platform (ITE) correlated significantly with various mention indicators, especially with PDF file page count in Google ($r=.75$) and total page count in Scholar ($r=.68$). However, a very low correlation was obtained with PDF page count in Google Scholar ($r=.31$), when it was precisely this indicator which should have been the most accurate in capturing the number of articles deposited in an institutional repository; it returned very low indexing ratios.
- With regard to the correlation of ITE with mention indicators, unexpectedly significant results were achieved with the number of URL mentions ($r=.63$), which demonstrates that despite the document noise of this indicator, the results do have certain value.
- Finally, almost no correlation was observed between ITE and indicators related to hyperlinks, both for the number of referring domains ($r=.26$) and for MzRank ($r=.22$).
- The PCA clearly shows the separation between performance in page count and visibility, and how the URL mention indicator seems closer to the page count than to the visibility indicators, when by their nature the opposite should be true.

2. DISCUSSION

Indexing ratios on Google & Google Scholar

The complete lack of correspondence between the repository records and the data provided by the Google & Google Scholar should be noted. Equally striking are the highly marked discrepancies in information between the search engines themselves: they only coincide in their extremely low indexing values for PDF documents.

This raises a preliminary question about the reliability and validity of the data search and recovery process ("site" command), the technical indexing mechanisms of the robots used by Google and Google Scholar and/or the deficient web architecture of the repositories themselves, which could well be the cause that lies behind the other aspects. Similarly, the design of the database of some of the repositories may prevent the accurate retrieval of indicators by search engine bots (a concept known as the invisible internet), although the development of applications such as DSpace (widely used in the installation of this study's sample repositories) has eliminated this problem.

With regard to Google (which should, in principle, index everything to achieve its goal of making the world's information universally accessible), the inordinately high page count data (well above real values) must be due to the counting of files that are not specifically items of the collection studied, i.e., files pertaining to the software itself or other information hosted by the server being analysed (easily verifiable by manually browsing through the results returned for the "site" query in the search engine).

Regarding the number of PDF documents, although exact figures for this document type in the repositories under study are not known, such a low indexing ratio is very strange. The pervasive use of the PDF format is an irrefutable fact in academia (Aguillo 2009), and it is very odd that academic repositories such as those studied here, which often contain scholarly output – theses, articles, reports and other academic documents (course syllabi, teaching materials) – have such a low percentage, save a few notable exceptions (<lume.ufrgs.br>, <repositorio.ufsc.br>). It is therefore plausible to conclude that Google underrepresents the scientific and academic content of the repositories.

By contrast, the total number of documents indexed in Google Scholar, contrary to Google, is well below what was expected. The low item indexing ratios in Google Scholar (whose database is not the same as Google's) are consistent with those obtained previously by Arlitsch and O'Brian (2012), who detected low indexing ratios in the United States for repository articles in Google Scholar, where only 30% of documents stored in the 21 repositories that formed their sample were included indexed in Google Scholar. Using the same methodology (the query: "site: repositoryURL") in this study, the indexing rate was only 34.2%.

Lower (17.1%) is the indexing ratio we found in June 2014 for the documents on the World Bank's Open Knowledge Repository (Martín-Martín et al, 2014).

In any case, there are several reasons that may explain why the overall data should be viewed with some caution. Aaron Tay has sharply summarized them in a recent post on his blog entitled "[8 surprising things I learnt about Google Scholar](#)"¹.

First, because the "site" operator does not return all the items that Google Scholar has indexed for a repository (special caution should be taken with URLs where the suffix PDF does not appear explicitly), which means it is not exhaustive. Second, because the system of grouping multiple versions of an article operates in such a way that one version is taken as the "primary" version. This process is done automatically, although authors may also manually select which is the main version of the article. The "site" command theoretically only returns data for the main version (though this not always happens). This means that if an article is hosted on different platforms (e.g. journal and repository), and if the primary version is the one published in the journal, the "site" operator applied to the repository will not count the item and vice versa, although it is indexed on both platforms.

This is the reason behind the fact that the indexation rate of the World Bank's Open Knowledge Repository has been much lower than the Latin-American repositories. This repository contains a large number of articles published in journals (over 10%) and related documents, whose versions appear and subsumed in other URLs.

Whenever the search strategy consists of a sample of papers searched individually, the indexing rate increases significantly. This corroborates the known issue of the underrepresentation of the number of results Google Scholar yields when a "site" command search is conducted. According to the study conducted by Doemeland and Trevino (2014), which uses the former methodology, almost 75% of the sample was indexed in Google Scholar. In our previous Digest ([Google Scholar Digest, n 2](#)) we already confirmed this².

This particularly affects the accuracy of Google Scholar in measuring the performance of repositories with this indicator, and largely explains the low values. It also opens up a future research line which should consider whether the repositories with better indexing ratios in Google Scholar are also those with higher numbers of primary versions amongst their items, which may explain the better results of some repositories compared to others.

What does stand to reason is that Google Scholar indexes far fewer PDF documents than Google, given the requirements and recommendations that this search engine provides institutional repository webmasters for indexing documents. These include the following³:

¹ http://musingsaboutlibrarianship.blogspot.sg/2014/06/8-surprising-things-i-learnt-about.html#.U6mZhpR_t3E

² <http://googlescholar Digest.blogspot.com.es/2014/06/world-banks-policy-reports-google-scholar.html>

³ <http://scholar.google.com/intl/en/scholar/inclusion.html>

- ❖ “If you’re a university repository, we recommend that you use the latest version of Eprints (eprints.org), Digital Commons (digitalcommons.bepress.com), or DSpace (dspace.org) software to host your papers.
- ❖ To be included, your website must make either the full text of the articles or their complete author-written abstracts freely available and easy to see when users click on your URLs in Google search results.
- ❖ Automatic crawlers need to be able to discover and fetch the URLs of all your articles, as well as to periodically refresh their content from your website. Browse interface is necessary for the search robots to discover the URLs of your articles. We recommend that the URL of every article is reachable from the homepage by following at most ten simple HTML links.
- ❖ Your website must not require users (or search robots) to sign in, install special software, accept disclaimers, dismiss popup or interstitial advertisements, click on links or buttons, or scroll down the page before they can read the entire abstract of the paper.
- ❖ Sites that show login pages, error pages, or bare bibliographic data without abstracts will not be considered for inclusion and may be removed from Google Scholar.
- ❖ Since Google refers users to your website to read the papers, your webpages must be available to both users and crawlers at all times. The search robots will visit your webpages periodically in order to pick up the updates, as well as to ensure that your URLs are still available. If the search robots are unable to fetch your webpages, e.g., due to server errors, misconfiguration, or an overly slow response from your website, then some or all of your articles could drop out of Google and Google Scholar.
- ❖ Your files need to be either in the HTML or in the PDF format. PDF files must have searchable text, i.e., you must be able to search for and find words in the document using Adobe Acrobat Reader.
- ❖ Each file must not exceed 5MB in size. To index larger files, or to index scanned images of pages that require OCR, please upload them to Google Book Search.”⁴

Arlitsch and O’Brian (2012), while noting the limitations of the “site” command, found that the main causes are the metadata schema used and the navigability and information architecture features, which do not help the search engine robots carry out the indexing processes correctly. Indeed, they applied various changes to the description schema (rejecting Dublin Core in favour of other schemas recommended by Google Scholar, such as Highwire Press), and then indexing ratios improved significantly over time.

These limitations of Google Scholar in measuring the presence of repository contents also contrast with the policy of certain products of this company, such as Google Scholar Metrics, which quantifies the scholarly impact of repositories (Delgado López-Cózar & Robinson-García 2012).

⁴ <http://scholar.google.com/intl/en/scholar/inclusion.html>

In short, it may be concluded that the low repository content indexing ratios are mainly due to these two limitations: the use of description schemas that are not compatible with Google Scholar (repository design) and the reliability of the web indicators (search engines).

Finally, it was found that the queries that combined the overall page count with the PDF file type in Google were those that achieved more optimal results, and that were most similar to the data that the repositories themselves indicated with regard to the size of their collections. This may have been determined by the fact that primary versions are not accounted for in the search – whereas in Scholar they are – which clearly underrepresents the presence of repositories when measured by the “site” command.

The final conclusions of this study highlight the insufficient dissemination of open access scholarly literature (crucially in terms of web visibility) in a medium (the Web) that is by definition its natural environment, and in a context (Latin America), in which scholarly production requires extra visibility because it lies outside the academic mainstream (i.e. not published in journals indexed in WoS or Scopus).

Given the weight of the green route in the dissemination of OA scholarly literature, and the importance of Google (and Google Scholar) to the search and use of academic information, the low visibility of the contents could significantly affect the real use of OA by end users. It would appear to be generating a great hidden mass of open access content, from institutional repositories, which neither Google, in the first instance, or users, in the last instance, can locate.

The lack of web visibility of the analysed repositories is determined by the low indexing ratios of their content (both in Google and Google Scholar), since a low web presence determines a corresponding low web visibility.

These low indexing ratios are, in turn, determined by the use of description schemas that are ill-suited to Google and inadequate web navigability, factors already outlined by Arlitsch and O’Brien (2012). Additionally, this study has also identified certain technical limitations in the use of web indicators in Google and Google Scholar to measure this indexing.

Therefore, we consider that neither Google nor Google Scholar are accurate or representative of the actual page count of open access content published by Latin American repositories; this may indicate the existence of a hidden, non-indexed side of OA.

In any case, the technical limitations of Google Scholar, in only counting primary versions of articles, tilt the balance towards the use of Google to measure page count, despite the fact that the document noise is greater. However, a thorough analysis of the real influence of the primary version search and accuracy of the “site” command in repository performance in Google Scholar (which requires an item by item analysis of each collection) is deemed necessary.

Much of the solution to these problems is purely technical, 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. This must include a rethinking of the goals that must be achieved to guarantee the success of a repository, for which presence and visibility in search engines must bear greater weight.

However, the results come from the analysis of a small sample of repositories, and should be widened in the future to larger samples in order to draw more definitive conclusions.

References

- Aguillo, Isidro F. (2009). Measuring the institutions' footprint in the web. *Library Hi Tech*, 27(4), 540-556.
- Arlitsch, K. & O'Brian, P.S. (2012). Invisible institutional repositories: addressing the low indexing ratios of IRs in Google. *Library Hi Tech*, 30(1), 60-81.
- Delgado López-Cózar, E. & Robinson-García, N. (2012). Repositories in Google Scholar Metrics: what is this document type doing in a place as such. *Cybermetrics*, v. 16.
Available at <http://cybermetrics.cindoc.csic.es/articles/v16i1p4.pdf> (accessed 15 March 2014).
- Doemeland, Doerte & Trevino, James (2014). Which World Bank reports are widely read? *World Bank Policy Research Working Paper*, n. 6851.
<https://openknowledge.worldbank.org/bitstream/handle/10986/18346/WPS6851.pdf?sequence=1>
- Martín-Martín, A.; Ayllón, J.M.; Orduña-Malea, E.; Delgado López-Cózar, E. (2014). The World Bank's policy reports in Google Scholar. Are they visible, cited, and downloaded?. *EC3 Google Scholar Digest Reviews*, n. 2.