Effectiveness of OpenAIRE, BASE, Recolecta and Google Scholar at finding Spanish articles in repositories

Francisca Abad-García (Corresponding author)
University of Valencia. History of Science and Documentation Department
Blasco Ibáñez, 17, 46010 Valencia (Spain)
Phone (+34) 963 864 950
Fax (+34) 393 864 091
e-mail: abad@uv.es

Aurora González-Teruel
University of Valencia. History of Science and Documentation Department
Blasco Ibáñez, 17, 46010 Valencia (Spain)
Phone (+34) 963 864 949
Fax (+34) 396 386 4091
e-mail: agonzal@uv.es

Javier González-Llinares
University of Valencia. History of Science and Documentation Department
Blasco Ibañez, 17, 46010 Valencia (Spain)
Phone (+34) 963 864 949
Fax (+34) 393 864 091
e-mail: jagonlli@alumni.uv.es

ABSTRACT

This paper pretends to explore the usefulness of OpenAIRE, BASE, Recolecta and Google Scholar (GS) for evaluating Open Access (OA) policies that demand deposit in a repository.

A case study was designed, focusing on 762 financed articles with a project of FIS-2012 of the Instituto de Salud Carlos III, the Spanish national health service's main management body for health research. Its finance is therefore subject to the Spanish Government OA mandate. A search was carried out for full-text OA copies of the 762 articles, using the four tools being evaluated and with identification of the repository housing these items.

Of the 762 articles concerned, 510 OA copies were found of 353 unique articles (46.3%) in 68 repositories. OA copies were found of 81.9% of the articles in PMC and copies of 49.5% of the articles in an institutional repository(IR).

BASE and GS identified 93.5% of the articles, OpenAIRE 86.7%. Recolecta identified just 62.2% of the articles deposited in a Spanish IR. BASE achieve the greatest success, by locating copies deposited in IR, while GS found those deposited in disciplinary repositories. None of the tools identified copies of all the articles, so they need to be used in a complementary way when evaluating OA policies.

INTRODUCTION

Many institutions currently operate policies that ask their researchers to disseminate their scientific production openly, whether by archiving their articles into a repository or publishing them in Open Access (OA) journals. In 2011, the Spanish government established a mandate requiring the obligatory deposit, in an institutional or thematic repository within 12 months of publication, of one copy of the final version of articles relating to publicly financed projects. Publication in an OA journal does not constitute any waving of this obligation (Ley 14/2011).

In 2007, and in order to make deposited content more visible, REBIUN (the network of Spanish university libraries) and FECYT (the Spanish foundation for science and technology) created Recolecta, a service provider that brings together OA items distributed via Spanish journals or repositories that comply with the guidelines of OAI-PMH (Open Archives Initiative-Protocol Metadata Harvesting) and the DRIVER (Digital Infrastructure Vision for European Research) directives. It currently includes 68 institutional repositories (IR)[1] of the 83 that exist in universities and research centers[2]. Recolecta has thus been added to a list of service providers[3] that includes two major European metadata harvesters (MH): OpenAIRE and BASE.

OpenAIRE (Open Access Infrastructure for Research in Europe) was created in 2009 to monitor compliance with the European Union's OA policy, although its scope has recently been expanded to cover further financing bodies (Rettberg & Schmidt, 2015). It covers more than 900 data providers, including European and non-European repositories, of which 58 are Spanish IR[4]. BASE (Bielefeld Academic Search Engine), established in 2004 by the University of Bielefeld, collects scientific content from more than 5,500 OAI-PMH-compliant sources (Lösch, 2011), of which 59 are Spanish IR.

Although MH are designed for finding copies deposited in the repositories covered by them, they are rarely used to verify the availability of OA copies of articles (Norris, Oppenheim & Rowland, 2008; Koskinen et al., 2010), given that MH does not cover copies placed in many other web locations whose identification is interesting for providing a global picture of the progress OA. For this purpose web search engines (SE) such as Google and Google Scholar (GS) are used in preference (Bjork, Roos, & Lauri, 2009; Bjork, et al., 2010; Archambault, et al., 2013; Khabsa & Giles, 2014; Jamali & Navabi, 2015), because their broad range of resources covered, albeit visibility limitations of IR content have been detected in them (Arlisch & O'Brien, 2012; Orduña-Malea & Delgado-López-Cózar, 2015).

Both types of tool therefore have their own advantages and disadvantages when it comes to finding documents deposited in a repository. In this context the object of the study has been to evaluate the effectiveness of OpenAIRE, BASE, Recolecta and GS as search tools for finding OA versions of articles that should have been deposited in a repository, for the purpose of verifying their usefulness when checking compliance with OA policies that impose this condition.

METHODS

A case study was designed, focusing on the Instituto de Salud Carlos III (ISCIII), the Spanish national health service's main management body for health research, whose calls for funding are subject to the Spanish Government OA mandate.

The 2012 call for funding of ISCIII's health research fund (FIS) was selected, and a search was made for articles published between 2012 and 2014 on the basis of projects financed by this call for funding, covering the main collection of the Web of Science (WOS) and using the search term "PI12/*" applied to the Grant Number (FG) and Funding Text (FT) fields. These characters correspond to the initial part of the code of each project grant awarded[5]. After manual verification of the results, a reference population of 762 articles was obtained.

A manual search of titles was carried out in OpenAIRE, BASE, Recolecta and GS to locate deposited full-text OA copies of the articles concerned. The name and corresponding type of repository were registered for each copy found: disciplinary (DR) and institutional (IR), with an indication of the Spanish items among the latter.

OpenAIRE, BASE y GS were evaluated for their efficiency in identifying copies of all articles deposited (globally and by type of repository), while Recolecta was evaluated with respect to copies deposited in Spanish IR. The degree of compliance with Spanish OA policy was expressed in percentage terms as a secondary result for the overall set of articles.

RESULTS AND DISCUSSION

Identification of OA articles

Of the 762 articles studied, 353 (46.3%) had at least one copy in a repository. This figure represents the percentage of compliance with the Spanish OA mandate for this population of articles.

Nevertheless, due to the study design it cannot be generalized to the entire population of Spanish publicly founded articles like it was the 35.3% obtained by Borrego (2016) from a sample of interdisciplinary studies. However this might constitute an indicative figure for the area of biomedicine, given that the FIS call for funding is the biggest budget item of the ISCIII.

Of the 353 deposited articles more than a copy was found for 115 (32.6%) (table 1). This circumstance increases the likelihood that at least a copy could be retrieved.

Table 1. Copies of articles deposited in repositories

No. of articles	No. of repositories	No. of copies
238	1	238
87	2	174
18	3	54
6	4	24
4	5	20
353		**510**

Deposit location

The 510 copies found were in 68 repositories, of which two were thematic and 66 institutional being 21 Spanish.

The DR housed 289 copies (without overlap between them) of 289 articles (81.9%); 288 in PubMed Central (PMC) and one at arXiv.org (Table 2). The PMC figures, which include PMC-Europe, PMC-USA and PMC-Canada, demonstrate its important role in the implementation of OA policies in biomedicine, given the corresponding document-incorporation policy. These PMC copies come from the deposit of the final, accepted manuscript articles of authors, and also from the direct deposit of articles published in OA journals, the delegated deposit of articles made OA by means the payment of a processing charge, and from the PMC Back Issue Digitalization Project (Europe PMC Consortium, 2014).

The IR plays a minor role in policy implementation in this case, as the copies housed in the 66 IR correspond to only 175 (49.6%) articles (Table 2). Of these, 134 (76.6%) were found in the 21 Spanish IR. This was to be expected, given the nature of the population studied.

Table 2. Copy location by retrieval tool

	No. of repositories			**No. of copies**			**No. of articles**		
	DR	**IR**	**Total**	**DR**	**IR**	**Total**	**DR**	**IR**	**Total**
Recolecta	0	19	19	0	92	92	0	83	83
OpenAIRE	2	41	43	268	144	412	268	123	307
BASE	1	56	57	273	190	463	273	159	331

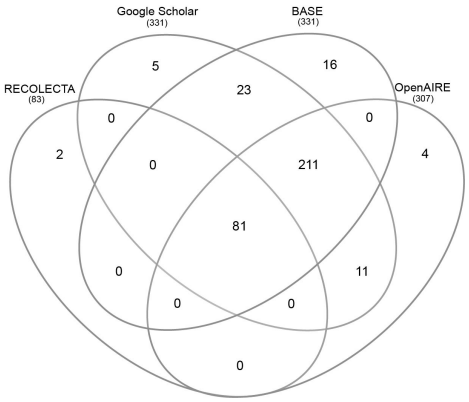
GS	1	42	43	286	165	451	286	150	331
Total unique items	2	66	68	289	231	510	289	175	353

Retrieval success

BASE and GS retrieved copies of 331 articles (93.5%), while OpenAIRE retrieved 307 (86.7%) (Table 2). None of the three tools identified all 353 articles, and they all identified articles in an exclusive manner; 16 in the case of BASE and five each for GS and OpenAIRE respectively (Fig. 1).

Recolecta retrieved copies of 83 (62.2%) of the 134 articles deposited in the 21 Spanish IR. BASE retrieved 124 (92.5%) and OpenAIRE 105 (78.4%). Only two of the articles detected by Recolecta were not retrieved by the other three tools.

Figure 1. Identification of OA articles by the tools evaluated



Retrieval success by repository type

The greatest success in terms of retrieval of copies deposited in DR corresponds to GS (99%), followed by BASE (94.5%) and OpenAIRE (77.8%) (Table 2). All three sources include PMC as a data provider. BASE uses PMC-USA, while OpenAIRE relies on PMC-Europe. It is assumed that the use of GS applies in all of these cases. Although PMC-Europe states that it retrieves open-access material via any of the PMC nodes, the OpenAIRE results indicate visibility problems with content obtained using this tool. On the other hand, the success of GS indicates improved content visibility for this repository with respect to the 25% attributed by Jacso (2008).

Whenever a copy was housed in an IR, BASE, the MH with greatest coverage, achieved the best success in terms of retrieval, by finding at least one copy of the 159 articles (90.9%) in 56 of the 66 IR identified.

GS only managed, despite its wide (but somewhat opaque) coverage of resources, to identify 85.1% of the articles with copies deposited in an IR, thereby demonstrating the previously mentioned difficulty with the indexing of such content (Arlisch & O'Brien, 2012).

OpenAIRE retrieved copies of 123 articles (70.2%) deposited in 41 of the 66 IR identified. This figure reflects the shortcomings of several repositories in terms of coverage prior to their more-recent expansion (Rettberg & Schmidt, 2015).

When considering in detail the retrieval of copies housed in Spanish IR, we see that Recolecta, which was specifically designed for this purpose, was only able to identify copies of 62.2% of the article deposited there, while BASE retrieved 92.9% and GS 78.4%. The shortcomings of Recolecta in terms of copy detection rates could be due to a lack of coverage on the part of the repositories concerned (it

identified copies in 19 of the 21 Spanish repositories), and also to the lack of visibility of the articles placed in the repositories theoretically offering this coverage. This nevertheless requires further research.

CONCLUSIONS

This study evaluates the efficiency of four tools used to retrieve copies of financed articles lodged in a repository. In spite of a limited ability, given the design of the study, to generalize the results obtained, it is evident that no single tool can exhaustively identify copies of studies kept in such repositories.

In the case of the evaluation of OA policies, the use of a single tool would give a distorted image of the actual situation; further exacerbated by a smaller evaluated population. This can apply to an institution, specific projects or research groups. In the case studied here, BASE and GS have proved to be the most effective tools, albeit complementary ones. BASE performed better when retrieving copies from IR, while GS was better at detecting copies deposited in PMC. Recolecta is on the other hand far from achieving the objective for which it was designed: making more visible Spanish OA science.

ACKNOWLEDGMENTS

This study was carried out under the project "Open Access to Science in Spain" (CSO2014-52830-P) of the Spanish R&D Plan funded by the Spanish Ministry of Science and Innovation.

REFERENCES

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Rebout, L., & Roberge, G. (2013). *Proportion of Open Access Peer-Reviewed Papers at the European and World Levels—2004-2011*. Science-Metrix. Retrieved from http://www.science-metrix.com/pdf/SM_EC_OA_Availability_2004-2011.pdf

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories. *Library Hi Tech*, *30*(1), 60–81. http://doi.org/10.1108/07378831211213210

Bjork, B.-C., Roos, A., & Lauri, M. (2009). Scientific Journal Publishing: Yearly Volume and Open Access Availability. *Information Research*, *14*(1). Retrieved from http://www.informationr.net/ir/14-1/paper391.html

Bjork, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLoS ONE*, *5*(6), e11273–9. http://doi.org/10.1371/journal.pone.0011273

Borrego, N. (2016). Measuring compliance with a Spanish Government open access mandate. *Journal of the Association for Information Science and Technology*, *67*(4), 757–764. http://doi.org/10.1002/asi.23422

Europe PMC Consortium. (2014). Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, gku1061.

Jacsó, P. (2008). Google Scholar revisited. *Online Information Review*, *32*(1), 102–114. http://doi.org/10.1108/14684520810866010

Jamali, H. R., & Nabavi, M. (2015). Open access and sources of full-text articles in Google Scholar in different subject fields. *Scientometrics*, *105*(3), 1635–1651. http://doi.org/10.1007/s11192-015-1642-2

Khabsa, M., & Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. *PLoS ONE*, *9*(5), e93949–6. http://doi.org/10.1371/journal.pone.0093949

Koskinen, K., Lappalainen, A., Liimatainen, T., Nevalainen, E., Niskala, A., & Salminen, P. J. (2010). The current state of Open Access to research articles from the University of Helsinki. *ScieCom*, *6*(4). Retrieved from http://journals.lub.lu.se/index.php/sciecominfo/article/download/4761/4332

Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación. *Boletín Oficial de Estado* nº131, 02/06/2011.

Lösch, M. (2011). A Multidisciplinary Search Engine for Scientific Open Access Documents. In R. Depping & S. Christiane (Eds.), *Elektronische Schriftenreihe der Universitäts- und Stadtbibliothek Köln* (Vol. 2, pp. 11–15).

Norris, M., Oppenheim, C., & Rowland, F. (2008). Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR. *Online Information Review*, *32*(6), 709–715.

Orduña-Malea, E., & Delgado López-Cózar, E. (2015). The dark side of open access in Google and Google Scholar: the case of Latin-American repositories. *Scientometrics*, *102*(1), 829–846. http://doi.org/10.1007/s11192-014-1369-5

Rettberg, N., & Schmidt, B. (2015). OpenAIRE. Supporting a European open access mandate. *College Research Libraries News*, *76*(6), 306–310.

---

[1] https://www.recolecta.fecyt.es/repositorios-recolectados

[2] http://www.accesoabierto.net/repositorio

[3] https://www.openarchives.org/service/listproviders.html

[4] https://www.base-search.net/about/en/about_sources_date.php?menu=2&submenu=1

[5] http://gesdoc.isciii.es/?action=download&id=05/11/2012-4cceaf17b0