Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Towards reliable data – counting the Finnish Open Access publications

Jyrki Ilva*

National Library of Finland, P.O.Box 26, 00014 University of Helsinki, Finland

Abstract

As in many other countries, the scholarly Open Access movement has been gaining momentum in Finland in recent years. There is a growing political pressure for taking steps towards a more comprehensive Open Access availability of Finnish research publications. One of the ideas currently under discussion is whether it would be possible to use Open Access availability of research publications as a factor in the funding model for the Finnish universities.

However, a key challenge is that at the moment there is no reliable, comprehensive data on how much of the Finnish research output actually is freely available. This paper discusses some of the plans on how to address this situation on a national level. The discussion may provide useful insights for other countries facing similar questions.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords: Open Access; Finland; research organizations; publication data; national data collection

1. Introduction

Although Open Access was for many years a relatively low-profile issue in Finland, the situation has been changing recently. This is partly due to Open Science and Research (http://openscience.fi/), a big national project funded by the

* Corresponding author. Tel.: +358-50-3182231.
E-mail address: jyrki.ilva@helsinki.fi

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Ministry of Education and Culture, which started out in 2014. In addition to research publications, the scope of the project encompass open research data and research methods.

A number of concrete steps towards Open Access have already been taken. The leading Finnish research funder, Academy of Finland, launched an Open Access mandate in 2015. In addition, a number of universities have their own Open Access mandates. Most of the Finnish research organizations are running institutional repositories, into which the researchers can self-archive Open Access copies of their publications.

The idea of using Open Access availability of research publications as a factor in the funding model for the Finnish universities has been under discussion for a couple of years. However, one of the issues that have been slowing down planning is that there is no reliable, comprehensive data on how much of the Finnish research output is already freely available. This paper discusses some of the plans on how to address this situation.

2. Collecting the data

The Ministry of Education and Culture has been collecting publication data from the Finnish universities since 2011. The quality of the data has great importance, as 13% of all state funding for the universities (more than 200 million euros a year) is distributed based on the number and quality of the publications. Since 2012 the Ministry has collected publication data from the universities of applied sciences as well. Data collection from the state research institutes and central hospital districts started in 2015.

The processes of data collection are currently under extensive development. CSC - IT Center for Science is responsible for the collection of the data, and it has been working in close co-operation with the research organizations, the Ministry of Education and Culture and the Finnish Federation of Learned Societies. The Federation of Learned Societies is running the Publication Forum, a national rating system of scientific publication channels, which is used in the university funding model for determining the quality of the publications. The combined publication data is available for viewing at the national Juuli research publications portal (http://www.juuli.fi), which has been developed by the National Library of Finland.[1]

So far the research organizations have reported their publications once a year, which has meant that the combined national data has been available only after a significant delay. However, starting from the year 2016 the process will be gradually automatized. Whenever possible, the publication data will be harvested by CSC from the CRIS of each organization and the records coming from different organizations will be de-duplicated and combined more or less in real time. The organizations can also easily update their records via the regular harvesting process. In addition, the combined and enriched national data will be available for the reporting organizations via APIs, and they can utilize it in their own systems as well. The national Juuli portal will show up-to-date information on publications once its processes have been upgraded by the end of this year.

Among the universities, the growing importance of national data collection has led to a widespread adoption of new CRIS platforms, which provide tools for the local data collection processes. Several of the universities use SoleCRIS (SoleNovo), a domestic product, but most of the recently implemented systems have been based on Pure (Elsevier) or Converis (Thomson Reuters).

However, as already mentioned, the quality of the data on which of the Finnish research publications are freely available is currently far from satisfactory. As long as the data is not reliable enough, there is no way Open Access could be used as a factor in the university funding model. On the other hand, it makes sense to collect the OA data in connection with the collection of other research publication data, as has been done in many other European countries. This provides obvious benefits for streamlining the data collection processes. With combined data it is also easier to generate statistics on the prevalence of Open Access in various publication categories, as has been done e.g in Denmark, which has recently launched the Danish Open Access Indicator in connection with the Danish National Research Database (http://www.forskningsdatabasen.dk/).[2]

In Finland the metadata format used in the national data collection has from the beginning included a field for indicating whether the publication is an Open Access publication or not. Unfortunately, the quality of the OA data has been relatively poor, mainly because of two reasons. The definitions and instructions used in the data collection have not been as clear that they should have been, and since the OA status of the publications has not had any financial effect for the participating organizations, they have not had motivation to invest proper resources in the gathering or verification of the data.[3]

As already noted, the publication data is usually collected from the CRIS of each reporting organization. As in many other countries, in Finland the CRIS and the institutional repository are usually two separate system infrastructures.[4, 5] From the point of view of data collection this situation is by no means optimal, although at some organizations there may be integrated processes in which the publications are first deposited into a CRIS and then transferred automatically into a repository. This means that the organizations may have data on the same publications in two different systems. In this case it may require some extra effort to verify which of the publications listed in the CRIS have actually been deposited into the repository. Although this can be done on a local level, it might be useful to consider whether the repositories should be better integrated with the national data collection.

In general, there are two basic methods for determining the Open Access status of publications. One is to check the status of each publication individually. If this is done in centralized fashion by the library or by the research administration, the work obviously requires quite a bit of resources.

The other method is to utilize outside data sources in an automatized identification process. For example, it is possible to identify most of the Gold OA journal articles by checking whether the publication channel is listed in the Directory of Open Access Journals (DOAJ). Although DOAJ does not contain all of the Gold OA publication channels - especially after the recent purge of about 3000 journals - it does include most of the OA journals actually used by the researchers.[6]

It is also fairly straightforward to check automatically whether the publications have been uploaded into the local institutional repository. On the other hand, at the moment there is no reliable method for identifying OA articles published in Hybrid OA journals or OA publications that have appeared as book chapters or in conference proceedings. In addition, if the publication has been uploaded into a non-local repository instead of the local one (which is often the case especially with co-authored publications), there might not be a way to get automatic notification of this.

The best results are achieved by combining these two approaches, although it may require some extra effort to create processes that minimize the amount of duplicated work. It would be helpful if there would be reliable international data sources for e.g. Hybrid OA publications or for publications that are freely available in repositories around the world. There are a number of international initiatives - including OpenAire, SHARE and CHORUS - which may prove to be useful in this regard, although there is still a lot of work ahead before they are reliable enough to be utilized for this purpose.

3. Improving the quality of the data

The Ministry of Education and Culture has already made an effort to improve the quality of the OA information collected from the research organizations. Starting from 2016, the definitions of different types of Open Access have been clarified, and the fields used for collecting this data have been updated.[7] The participating institutions may need to update their technical systems to support the required fields, but since the OA data may be used for determining funding levels at some point, the institutions should be motivated enough to comply with the requirements. As the funding of the universities is based on data from the previous three years, the earliest possible starting date for using the OA data for funding purposes is 2019.

After some discussion it was decided that instead of using one OA status field in the national data collection with Gold and Green OA as separate options, it would be better have two separate fields, one indicating whether the OA publication has been issued in either Gold or Hybrid OA publication channel, and the other indicating whether the publication has been deposited into an open repository. This was deemed necessary as the same publication can be both issued in a Gold or Hybrid OA journal and deposited into a repository.

The organizations are also required to provide URL for each of the open access versions so that the OA status of the publication can be verified. For long-term availability purposes it is recommended that the URLs used in the reporting should be based on permanent identifiers. In the case of Gold or Hybrid OA publications - especially journal articles - this usually means using DOIs, while the Finnish repositories generally use URNs or Handles as a basis for the permanent addresses.

One of the key issues in OA monitoring is having a clear definition on which kind of publications count as OA publications and which don't. The Ministry of Education and Culture has tried to come up with a fairly strict definition of OA, which would still be permissive enough to be acceptable for all fields of research. The basic rules are the following:

- The publication can be read online in full and without restriction, printed out and copied at least for research use
- The publication is available either a) directly from the publisher's service or b) no later than after the end of an embargo period set by the publisher, through archiving in a repository dedicated to a specific organization or field of science
- The publication is freely available from a service provided by either a publisher or research organization that enables the harvesting of the metadata of publications and the indexing of them to other search services, and which supports citations and links to publications preferably using URLs based on persistent identifiers (DOI, URN, Handle)
- The freely available version of the publication is, depending on the publication contract or publisher's policy, either the author's last self-archived (peer-reviewed) version or the final version published via the publisher's service

Although some of the OA advocates stress that all OA publications should be published under a CC-BY license ("libre OA") so that their content can be re-used in data mining, there is no universal agreement on this.[8] Therefore it was decided that at least at this point it is enough that the publications are freely available for reading and printing ("gratis OA"). On the other hand, there were reasons why it was necessary to restrict types of the services which qualify as open access repositories. Although sites like Academia.edu or ResearchGate are very popular among the researchers, they are for-profit services which won't necessarily provide permanent Open Access to the publications or don't allow the harvesting of the metadata of the publications to other services.[9]

There are a number of other tricky questions mostly related to the publisher policies which may require further examination. One of them concerns embargoes and delayed OA. Although the embargo periods imposed by different publishers and journals vary a lot, the guidelines provided by the Ministry don't currently define a maximum length for the embargoes. This is something that may change later on. On the other hand, following the general international guidelines, delayed OA (i.e. articles published in journals which make all of their content freely available after a delay) does not count as OA, even if the delay used by some journals may be significantly shorter than the Green OA embargoes required by some of the other publication channels.[10]

Another issue which has generated discussion is which version of the publication should be deposited into a repository. According to the Ministry guidelines, the self-archived version of the publication should be author's last peer-reviewed version (of course, assuming the final publisher's version cannot be self-archived). In practice this means that the pre-prints deposited into e.g. ArXiv are not considered OA publications unless the author(s) later add a version that has gone through peer-review to ArXiv.

4. Potential future needs

Although the main objective of the Finnish OA monitoring plans is to provide reliable information on the share of research publications that are freely available, there are a number of other future directions that may also be worth exploring. Some of these would be easy to implement with the systems that are already under development, some of them would require more planning.

In addition to actual Open Access availability it would be possible to measure the potential for Green OA, as has already been done in Denmark as part of the Danish Open Access Indicator.[2] This would be relatively easy to do in Finland as well, as the Sherpa/Romeo information on the self-archiving policies of each journal has already been integrated into the national publication channel database, which is used as an information source both in the national data collection process and at the local level at some universities. Although the color codes used by Sherpa/Romeo are sometimes more cryptic and less reliable than one would hope, the integration would be very useful for both political and practical purposes.

One of the potential new needs that the national data collection might be expected to serve at some point is the monitoring of the Open Access fees (mostly Article Processing Charges) paid to the publishers. However, unlike some other European countries, Finland doesn't currently have centralized open access funds or institutional processes for the management of these payments. This means that at this point the research organizations wouldn't have been able to supply the information without a great deal of extra effort.[11]

It is still unclear whether the current research information systems will be used to monitor the flow of money associated with publication fees or whether there will be other systems and processes to take care of this. Of course, it would be beneficial if all of this information could be managed within the same system as all of the other data related to research publications, but it remains to be seen whether the current research information systems will be up to this new task at some point.

In addition, the OA monitoring information is also crucial for the negotiation of large-scale offsetting deals with the publishers, as proposed in a Max Planck Institute White Paper last year.[12] The idea of an offsetting deal is to use (some of) the money currently used for the Big Deal site licenses of restricted-access content for providing Open Access to the publications produced by the researchers affiliated with the organization. In Finland FinELib, the national licensing consortium of the libraries, is planning to add offsetting elements into its licensing deals with the major international publishers.

For this purpose it would be very useful to get a grip of all of the costs associated with the research publications, including the money spent on licensing or offsetting deals and the money spent on APCs.[13] This may be something that is out-of-scope for the current CRIS platforms and for the national data collection as well, especially as the CRIS is often managed by the research administration while the licensing deals are negotiated by the libraries. There would be obvious synergies and benefits in combining these issues, but it remains to be seen whether it will be feasible to develop systems that will be able to take care all of this, or whether it would make more sense to build separate systems which will share information between each other.

References

1. Ilva J. Juuli - a national portal for publication data. CSC News 3/2013, pp. 8-10. Retrieved May 19, 2016, from http://urn.fi/URN:NBN:fi-fe201402121466
2. Elbæk M. The Danish Open Access Indicator. *PASTEUR4OA Final Conference: Green Light for Open Access.* May 17-18, 2016, Amsterdam. Retrieved May 19, 2016, from http://www.pasteur4oa.eu/sites/pasteur4oa/files/generic/DanishOAindicator_Pasteur4OA.pdf.
3. Ilva J. OKM:n tiedonkeruun avoin saatavuus tieto. Report [in Finnish], March 2, 2015. Retrieved May 19, 2016, from http://urn.fi/URN:NBN:fi-fe2015090111081.

4. De Castro P, Shearer K, Summann F. The gradual merging of repository and CRIS solutions to meet institutional research information management requirements. "Managing Data-Intensive Science: the Role of Research Information Systems in Realising the Digital Agenda": *Proceedings of the 12th International Conference on Current Research Information Systems (2014)*. Procedia Computer Science 33: 39-46 (2014). Retrieved May 20, 2016, from http://hdl.handle.net/11366/197.
5. Ilva J. Integrating CRIS and repository - an overview of the situation in Finland and in three other Nordic countries. Presentation at *Open Repositories 2014*, June 9-13, 2014, Helsinki. Retrieved May 20, 2016, from http://urn.fi/URN:NBN:fi-fe2014070432242.
6. Baker M. Open-access index delists thousands of journals. *Nature*, 09 May 2016. Retrieved May 19, 2016, from http://www.nature.com/news/open-access-index-delists-thousands-of-journals-1.19871.
7. Publication data collection instructions for researchers 2016. English version. Ministry of Education and Culture, May 4, 2016. Retrieved May 19, 2016, from https://confluence.csc.fi/display/suorat/Julkaisutiedonkeruun+tutkijaohjeistukset.
8. Anderson R. CC BY and its discontents - a growing challenge for Open Access. *Library Journal*, Feb 19, 2015. Retrieved May 19, 2016, from http://lj.libraryjournal.com/2015/02/opinion/peer-to-peer-review/cc-by-and-its-discontents-a-growing-challenge-for-open-access-peer-to-peer-review/.
9. Matthews D. Do academic social networks share academics' interests? *Times Higher Education*, April 7, 2016. Retrieved May 19, 2016, from https://www.timeshighereducation.com/features/do-academic-social-networks-share-academics-interests.
10. Laakso M, Björk B-C. Delayed open access: An overlooked high-impact category of openly available scientific literature. *Journal of the American Society for Information Science and Technology*. 64 (2013): 7. pp. 1323-1329. Retrieved May 19, 2016, from http://dx.doi.org/10.1002/asi.22856.
11. Naukkarinen, P. Avoimen julkaisemisen tuen malli. Report [in Finnish], March 16, 2016. Retrieved May 19, 2016, from http://urn.fi/URN:NBN:fi-fe201603308918.
12. Schimmer, R, Geschuhn, K. K., & Vogler, A. (2015). Disrupting the subscription journals' business model for the necessary large-scale transformation to open access. *A Max Planck digital library Open Access policy white paper*. Retrieved April 30, 2015, from http://dx.doi.org/doi:10.17617/1.3.
13. Pinfield S, Salter J, & Bath PA. The 'total cost of publication' in a hybrid open-access environment: Institutional approaches to funding journal article-processing charges in combination with subscriptions. *Journal of the Association for Information Science and Technology,* Feb 13, 2015. Retrieved May 19, 2016, from http://dx.doi.org/10.1002/asi.23446