My repository is being aggregated: a blessing or a curse?

Petr Knoth, Lucas Anastasiou, Samuel Pearce
Knowledge Media institute
The Open University
United Kingdom

Abstract. Usage statistics are frequently used by repositories to justify their value to the management who decide about the funding to support the repository infrastructure. Another reason for collecting usage statistics at repositories is the increased use of webometrics in the process of assessing the impact of publications and researchers. Consequently, one of the worries repositories sometimes have about their content being aggregated is that they feel aggregations have a detrimental effect on the accuracy of statistics they collect. They believe that this potential decrease in reported usage can negatively influence the funding provided by their own institutions. This raises the fundamental question of whether repositories should allow aggregators to harvest their metadata and content. In this paper, we discuss the benefits of allowing content aggregations harvest repository content and investigate how to overcome the drawbacks.

The purpose of repositories and the need for aggregations

Requests from repositories that try to avoid being harvested go clearly against the main principle of why repositories have been established as documented in the SPARC's position paper on institutional repositories (Crow, 2002) - *the primary goal of repositories is to open and disseminate research outputs to a worldwide audience*.

The SPARC's position paper specifically says:

"For the repository to provide access to the broader research community, users outside the university must be able to find and retrieve information from the repository. Therefore, institutional repository systems must be able to support interoperability in order to ***provide access via multiple search engines and other discovery tools****. An institution does not necessarily need to implement searching and indexing functionality to satisfy this demand: it could simply maintain and expose metadata,* ***allowing other services to harvest and search the content****. This simplicity lowers the barrier to repository operation for many institutions, as it only requires a file system to hold the content and the ability to create and share metadata with external systems."*

In summary, to fulfill their main purpose, repositories must allow external systems to harvest both the metadata and the content. Without the possibility of aggregating repository content, it would also not be possible to realise the vision foreseen by the Confederation of Open Access Repositories (COAR), which states:

*"****Each individual repository is of limited value for research****: the real power of Open Access lies in the possibility of connecting and tying together repositories, which is why we need interoperability. In order to create a seamless layer of content through connected repositories from around the world, Open Access relies on interoperability, the ability for systems to communicate with each other and pass information back and forth in a usable format. Interoperability allows us to exploit today's computational power so that we can* ***aggregate, data mine, create new tools and services, and generate new knowledge from repository content.'****" [COAR, 2011]*

Is the open access market held back by protectionism?

At the moment, the approach of many repositories is that they are designed to be friendly to a predefined restricted set of (typically commercial) aggregators, in particular Google, while they are shut to other third-party systems. This creates an almost paranoid situation in which many open repositories are more open to closed

commercial systems than to the systems developed by the open access community itself.

For example, the open access aggregator CORE[1] has this experience with the French repository Archimer[2], which is registered in OpenDOAR[3]. Last year, the Archimer repository manager asked the CORE team not to harvest their content on the grounds of protecting their own repository usage statistics. This is regardless of the fact that, as discussed in the next section, an entirely opposite outcome might actually be the result.

While the CORE team has never received a similar request from a UK repository, which are the primary CORE's target user group, this situation occasionally occurs with publishers and less frequently with repositories. Many of the publisher systems provide content under some form of Creative Commons license allowing redistribution, such as CC-BY. These publishers are sometimes even proactively registered in DOAJ. For example, the request not to harvest content has been made in the past by the Eurasia journals that were registered in DOAJ and whose license allowed redistribution, more specifically the content was licensed as CC-BY-SA. We have also recorded a similar request from the OTHES repository based at the University of Vienna, despite the university staff referring to the content as open access and the repository being registered in OpenDOAR.

Restrictions on machine access to open access content seem to be a common practice in subject-based repositories (such as PubMed or Europe Pubmed[4]) as well as the systems of major commercial publishers. In these situations, however, the reasons for not allowing machine access to open access content are (at least to the public) unknown.

This kind of protectionism is not only unethical and disadvantageous for the scholarly community and the public, Groom (2004) even suggests it might be illegal as it, among other things, triggers concerns of unfair competition. It is critical that the community recognises that the open access movement cannot be truly successful unless the infrastructure market is liberated from abusive/monopolistic practices.

Can aggregations add value?

The experience of the team around the CORE aggregator actually shows that the majority of repositories are aware of the benefits aggregations bring. This is evidenced by the fact that over the last three years we have received far more opt-in requests than opt-out. Overall, the main role of aggregations is to fulfill and support use cases, which cannot be satisfied by individual repositories. These use cases include a) harmonised programmable access to all available OA metadata and content, which is needed to enable content re-use by new services possibly utilising text-mining, b) transaction access allowing users to explore content across repositories leading to an increased content visibility, for example, through cross-repository content recommendation or search and discovery tools, and c) analytical access to information allowing the monitoring of content growth, new trends in scientific disciplines, etc (see (Knoth, 2012, 2013) for more details). Thus, aggregations should functionally complement repositories creating a mutually beneficial ecosystem.

Can everybody win?

Coming back to usage statistics, we believe that even without the existence of aggregations, the OA philosophy dictates that the existence of multiple copies of an article on the Internet is not only permitted, but even

---

[1] http://core.kmi.open.ac.uk/search

[2] http://archimer.ifremer.fr/

[3] http://www.opendoar.org/

[4] Both sites restrict access to the OA content to machines by predefining a list of agents that can harvest content through the website. Agents that are not in this list are disallowed access using the Robots Exclusion Protocol. There is no procedure or a set of criteria for getting added to this list. Our question about what an agent needs to do to be added was left unanswered, suggesting the inclusion to the list is a result of some sort of protectionism. Nevertheless, Boonk (2005) argues that the Robots Exclusion Protocol does not have a legal status anyway, thus not respecting it cannot be legally enforced.

beneficial. Multiple copies allow for better preservation[5], lower network latency, might increase visibility, provide high re-use opportunities and keep the scholarly market free from monopoly. In addition, researchers seem to like copying of articles. Articles can thus be found on personal web pages or in systems, such as Mendeley or ResearchGate. Spreading of multiple copies of OA content should therefore be seen as something positive. Consequently, there is a need for a truly distributed approach for acquiring OA usage statistics.

In understanding of this context, we have realised that the whole issue of lost usage statistics can be resolved by attributing the usage from external systems (including aggregators) to the repositories of origin. We have tested this approach with the IRUS-UK[6] service, which is available in the UK, however the same approach can be applied universally.

The implementation

The IRUS-UK software has piloted an authoritative service for benchmarking download statistics from repositories. Once the IRUS-UK software plugin is installed into a repository system, it informs the central IRUS-UK server whenever there is a download request from the repository. The IRUS-UK server is equipped with a smart filtering mechanism to percolate all robots and crawler requests, producing clean comparative statistics for repositories.

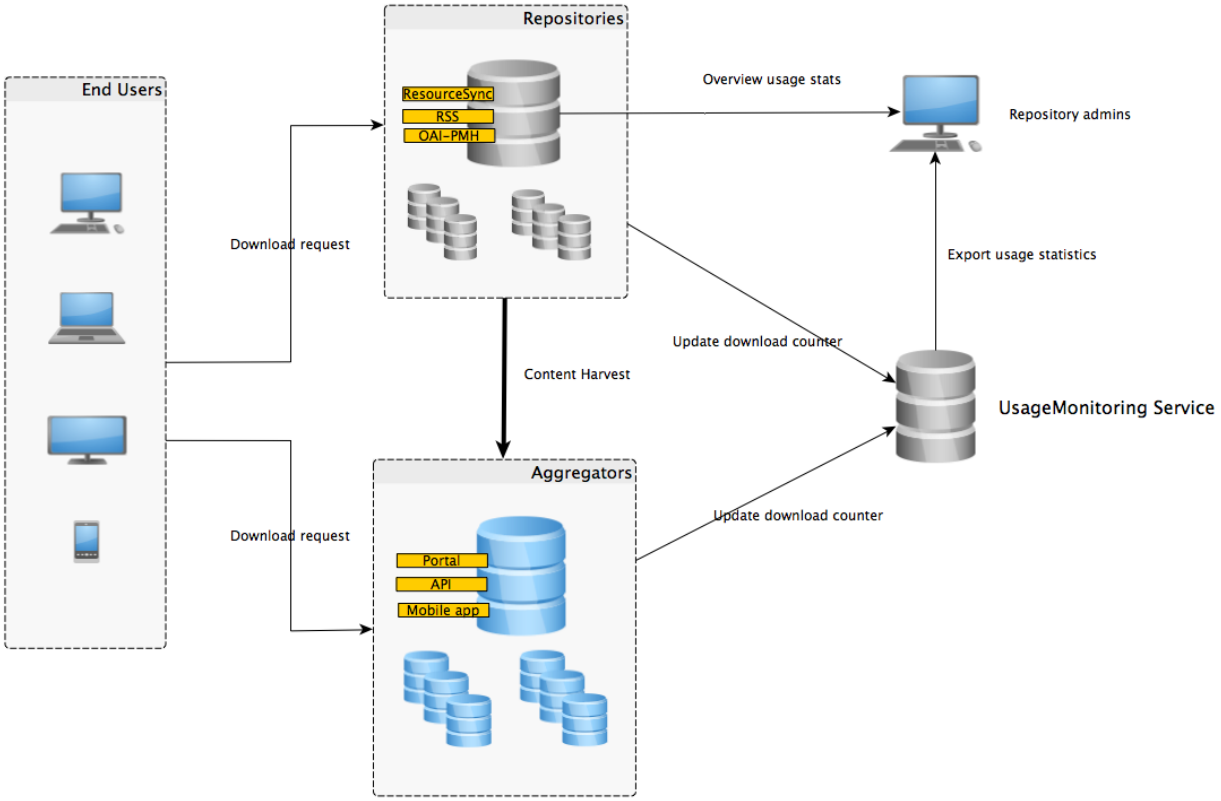

Figure 1: A scenario in which article downloads are attributed to the repository of origin even if the download is requested from an aggregator.

Our proposition is that the IRUS-UK software or a similar benchmarking can be installed also in aggregators. Since aggregators know the origin of the item, the benchmarking statistics can be adjusted on the IRUS-UK

[5] See, for example, the LOCKSS Program (Lots of Copies Keep Stuff Safe)

[6] http://www.irus.mimas.ac.uk/

server so that the repositories do not lose any information about the usage of the repository items.

CORE informs IRUS-UK usage monitoring service about every article download request by providing the article's OAI identifier plus some additional metadata. The OAI identifier itself is sufficient for determining the article's provenance. Thus, IRUS-UK uses OAI to identify the repository of origin of the downloaded publication. The download can then be attributed to repository statistics of the repository of origin.

The download request is not delayed by the usage monitoring service update as it is issued in asynchronous fashion and there is no dependency of the usage monitoring service update to succeed before serving. In the case of a network problem, usage updates are stored in a temporary buffer and pushed to the IRUS-UK service as soon as it is available, ensuring all download requests are tracked. (*no danger of non-tracked traffic because monitor service went down).*

Conclusion

This paper discussed the paradox of repositories needing to disseminate open access content to a worldwide audience, while keeping the repository usage statistics high to justify funding. Technically speaking, open repositories must provide unrestricted access to both commercial and not-for-profit aggregators allowing them to effectively harvest their content. We have experienced a few short-sighted approaches to tackle this problem in the past.

We propose a solution to this issue that satisfies both parties and consequently benefits the ecosystem of repositories and aggregators. This approach has been implemented in the UK by integrating IRUS-UK usage statistics service with the CORE aggregator. Even though the IRUS-UK service is currently available only for UK repositories, the approach can be extended worldwide.

References

[Crow, 2002] Crow, R. (2002). The case for institutional repositories: a SPARC position paper. ARL Bimonthly Report 223

[Bonk, 2005] Boonk, M. L., de Groot, D. R. A., Brazier, F. M. T., & Oskamp, A. (2005). Agent Exclusion on Websites. *LEA*, 13-20.

[Groom, 2004] Groom, J. (2004). Are 'Agent' Exclusion Clauses a Legitimate Application of the EU Database Directive?. SCRIPT-ed, 1(1).

[Knoth & Zdrahal, 2012] Knoth, P. and Zdrahal, Z. (2012) CORE: Three Access Levels to Underpin Open Access, D-Lib Magazine, 18, 11/12, Corporation for National Research Initiatives, http://dx.doi.org/10.1045/november2012-knoth

[Knoth, 2013] Knoth, P. (2013) CORE: Aggregation Use Cases for Open Access, Demo at Joint Conference on Digital Libraries (JCDL 2013), Indianapolis, Indiana, United States

[Rodriguez, 2011] Eloy Rodrigues and Abby Clobridge. 2011. The case for interoperability for open access repositories. Working Group 2: Repository Interoperability. Confederation of Open Access Repositories (COAR).