Collecting Inclusive Usage Metrics for Open Access Publications: the HIRMEOS Project

Javier Arias

Collecting Inclusive Usage Metrics for Open Access Publications: the HIRMEOS Project

Javier Arias

Introduction

1 Usage metrics have been widely adopted in Open Access works as an indication of the popularity or acceptance of a particular publication (Kumar, 2015). Inevitably, performance assessment and funding allocation are being based on these statistics (Giménez-Toledo *et al.* 2015), making metrics collection and reporting a fundamental need for any organisation producing and/or hosting digital monographs.

2 A main characteristic of Open Access publications is the ease of redistribution that their licences entitle, allowing dissemination across multiple platforms (Suber, 2007). However, while most distributing platforms produce usage metrics for their own analysis, these are rarely made available publicly via APIs that enabled programmatic collection of such metrics, preventing less technologically advanced organisations from harvesting this data.

3 Publishers who want to obtain metrics from distributing platforms struggle to find consistency in the book identifiers used, varying from ISBN numbers to URLs, DOIs, or even the publication title, or the platform's own database serials. Distributing platforms also struggle to obtain comparable data from peer organisations and to produce reliable metrics themselves. These intricacies mean a high entry barrier for small publishers and university presses that can neither implement an existing standard due to high validation fees nor can they develop their own metrics software.

4 In order to help tackling these problems, the HIRMEOS project presents Work Package 6, in which Open Book Publishers is working to create and populate a database of title-specific usage data-aggregating metrics across multiple different platforms and formats. Drivers are being developed to harvest alternative hosting platforms for usage data, and

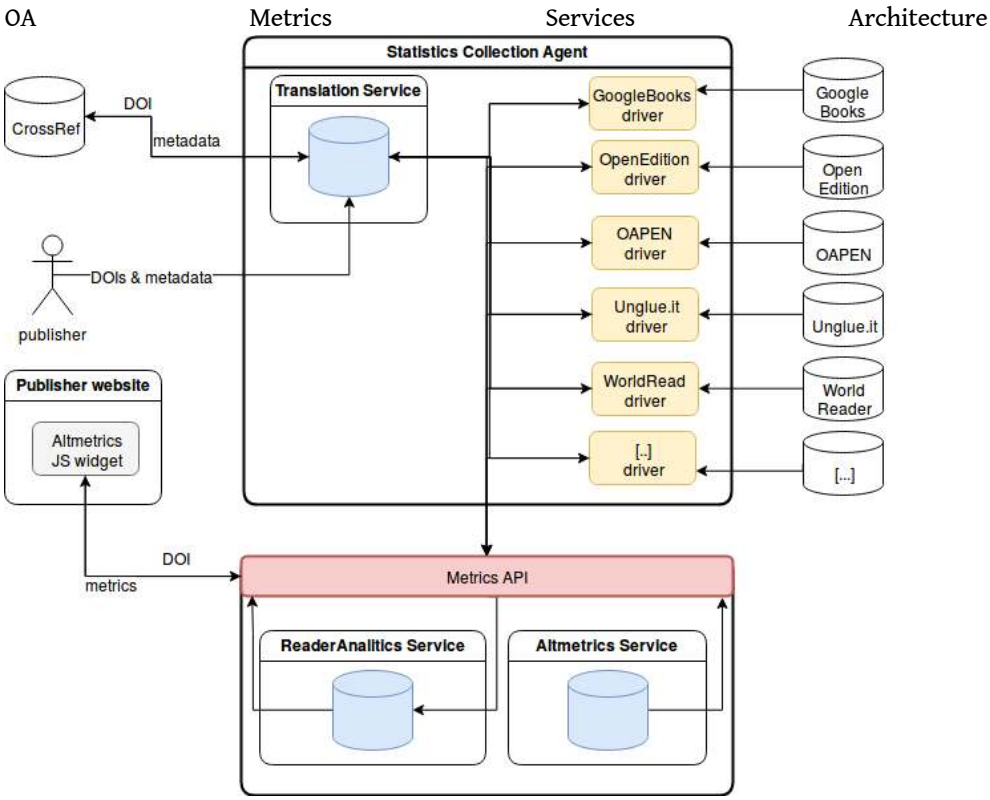
an API created for publishers and platforms to query the database and present the aggregate data on their own websites. All the code and architecture created for this project will be Open and be made available for others to freely download, adopt or adapt as they wish, facilitating broader uptake, collection, and presentation of this data for Open Access publications.

Design and Methodology

5 Collecting usage data for books presents a set of challenges different than for articles. Authors and publishers are interested in obtaining overall usage data for both the book and individual chapters within a book. Even citation identifiers such as the DOI remain limited for books. Books and individual chapters have different identifiers. Many platforms hosting digital editions of OA titles assign their own DOIs or permanent URL references to the content, as well as numerous different digital formats for ebooks exist and circulate.

Statistics Collection Agent

6 The Statistics Collection Agent will be developed in a modular fashion, where each of the modules constituting the software will be called drivers; one for each of the disseminating platforms which one wish to collect metrics from. These drivers will work independently of one another, minimising the risk of failure, and allowing future expansions and the implementation of new drivers; as well as publishers will not need to run or even install, drivers for those platforms in which they do not possess any content. The role of these modules will be both the download of metrics from its disseminating platform and the upload to the remote database via the metrics API. Publishers will be encouraged to schedule drivers to run daily via cron jobs in order to keep metrics updated. The list of drivers -subject to future expansion- that will be developed in this project includes Google Analytics, Access Logs, Google Books, Open Edition, OAPEN, Wikimedia, Unglue.it, The Classics Library, OpenAIRE, IRUS-UK, JSTOR, Matomo (Piwik), and World Reader. The Google Analytics and Matomo drivers will be provided as a generic template that will allow disseminating websites (and publishers who also act as such) to develop a driver for their own online readers; those who do not wish to share their usage data with commercial third-parties will be able to use the Access Logs driver to achieve the same purpose. An Access Logs based driver to provide download metrics will also be provided. Some of the disseminating platforms provide APIs that can be queried to obtain metrics, others require web scraping. Even though generic drivers for each use case cannot be provided given that each API or web interface will differ from one another, it is expected that publishers will be able to create new modules for platforms not covered in this project based on existing drivers that follow the same methodology they intend to produce. For instance, Google Books does not provide a metrics API, only a reporting interface where publishers and authors can download a CSV report; taking its driver as an example it would be trivial to develop a new module for a different platform which lacked an API but provided metrics reports. The architecture of the service can be appreciated in greater detail in the figure below.



Open Book Publishers: CC-BY

Identifier Translation Service

7 Each disseminating platform makes use of a different identifier for their content: Google Books using ISBN numbers, World Reader book titles, Open Edition URLs, etc. In order to produce reliable metrics for a publication, one must first convert the primary keys obtained from each platform to a common identifier.

8 The Identifier Translation Service consists of a database populated with identifiers for each publication the publisher or platform is interested in, stored in the form of Uniform Resource Identifiers (URI). URIs provide a simple and extensible means for identifying a resource (Berners-Lee *et al.* 2005), allowing the storage of any possible identifier a book may have, standardised using an Open Standard freely available to all. Thus, an ISBN number such as 978-1-78374-368-1 would be stored using the URN scheme (Moats, 1997) as urn:isbn:9781783743681; a DOI like 10.11647/obp.0001 using the "info" scheme (Van de Sompel *et al.* 2006) as info:doi:10.11647/obp.0001; and custom platform-specific identifiers may be stored using the "tag" scheme (Kindberg and Hawke, 2005).

9 Users may query this API providing either a URI or a title, and indicating the URI scheme and namespace of the identifiers they want to retrieve. Since identifiers are normally shared between the book and its chapters, the user may also specify the type of work they are interested in (e.g. monograph, book-section, book-set, etc.), and indicate whether they want a unique result to be retrieved by setting a strict boolean flag to true, which will trigger an error if multiple results are found in the query. Resolution of fuzzy titles to URIs is also possible using a result scoring system that uses Levenshtein distance to determine the fittest candidate in a search query.

10 There will be occasions where a single publication has been assigned multiple identifiers of the same type (e.g. multiple DOIs) and the translation service will fail to programmatically return a single mapping. In order to avoid heuristic resolutions, which are very error prompting, publishers will be allowed to assign a canonical URI within each scheme for each publication; thus ensuring a conscious resolution of the conflict. The translation API will, therefore, enable the storage of book metadata, and facilitate DOI conflict resolution through canonicalisation. Publishers will be able to perform these actions with the provided front end, or they can otherwise integrate them into their existing administration systems.

Measures in the Metrics API

11 The metrics database and API will enable the storage of results provided by the other services, where metrics will be stored against a URI representing the object of study, following the same specifications as the Translation Service.

12 Although the list could potentially be extremely extensive, disseminating platforms tend to keep the number of measures offered to a minimum, normally exposing only book views, but sometimes also page views, book downloads, and/or unique book views. One of the ambitions of defining this format is to achieve a sufficient level of flexibility that will allow its application to metrics collected from different platforms that are already using different criteria but also to anticipate to future specifications. Consequently, there cannot exist a predefined list of available measures in this standard, but rather a set of rules that allow defining them.

13 The concept of a measure is made of three components: a platform, a namespace, and a type; and that each measure will be identified by a unique name within its namespace. The platform component will identify where the data was collected from (e.g. Google Books, Open Edition, World Reader, etc.), and the type will reflect what the measure represents (e.g., a visit, a download, a session, etc.). These two components alone are actually sufficient to define a measurement, however, they fail to identify those measures that can be aggregated together. Instead of imposing a unique point of view, it is desired to accommodate each different interpretation of measure aggregation allowing the users of this standard to create their own definitions.

14 Each set of measure definitions will be collected using the “tag” URI scheme, under a particular namespace that indicates the organisation defining such measures (e.g. operas.eu), and will represent a standard of the different measures that can be aggregated together, as agreed by the parties involved in the definition of that particular namespace. Consequently, namespaces represent not only a flag for technical purposes but also an indication of the organisation that takes responsibility for its definition. For instance, partners of the OPERAS consortium may use the “tag:operas.eu,2018:readership” namespace to identify measures related to book readership that they consider to be comparable, differentiating book visits from Google gathered in Google Books using “tag:operas.eu,2018:readership:google-books”, from book downloads harvested from the OAPEN library “tag:operas.eu,2018:downloads:oapen”.

15 Although the flexibility of measure definition may not solve the lack of consistency in usage metrics, one of the main aims of this design is to raise awareness that we must not impose the use of metrics standards to organisations that may not have the resources to

implement them. However, it is also fundamental that the variety of statistics collected by these services is acknowledged when reporting them to users, and that a transparent and comprehensive broken down view is provided at all times.

Conclusion

16 Very few efforts have been made towards achieving a comprehensive and transparent mechanism to collect and aggregate usage metrics. Most platforms are limited to collecting their own usage, while many do not even comprehend the complexity of the matter. The HIRMEOS project poses a groundbreaking approach that enables metrics collection and aggregation from third-party platforms, which is currently a manual job for many scholarly publishers that lack the funds to find a technical solution to the problem.

BIBLIOGRAPHY

References

Berners-Lee, T., Fielding, R., & L. Masinter (2005). "Uniform Resource Identifier (URI): Generic Syntax". STD 66, RFC 3986, DOI 10.17487/RFC3986 https://www.rfc-editor.org/info/rfc3986

Giménez-Toledo, E., Mañana-Rodríguez, J., Engels, T., Ingwersen, P., Polonen, J., Sivertsen, G., Zuccala, A. A. (2015). "The Evaluation of Scholarly Books as Research Output. Current Developments in Europe." *Proceedings of the 15th International Society for Scientometrics and Informetrics Conference*, 469–476.

Kindberg, T., & Hawke S. (2005). "The 'tag' URI Scheme". RFC 4151, DOI 10.17487/RFC4151, https://www.rfc-editor.org/info/rfc4151

Kumar, A. (2015). *Research Evaluation Metrics*. DOI 10.5530/jscires.5.1.12, http://unesdoc.unesco.org/images/0023/002322/232210E.pdf

Moats, R. (1997). "URN Syntax". RFC 2141, DOI 10.17487/RFC2141, https://www.rfc-editor.org/info/rfc2141

Suber, P. (2007). "Open Access Overview". https://legacy.earlham.edu/~peters/fos/overview.htm

Van de Sompel, H., Hammond, T., Neylon, E., & Weibel S. (2006). "The 'info' URI Scheme for Information Assets with Identifiers in Public Namespaces." RFC 4452, DOI 10.17487/RFC4452, https://www.rfc-editor.org/info/rfc4452

ABSTRACTS

Open Access has matured for journals, but its uptake in the book market is still delayed, despite the fact that books continue to be the leading publishing format for social sciences and humanities. The 30-months EU-funded project HIRMEOS (High Integration of Research Monographs in the European Open Science infrastructure) tackles the main obstacles of the full integration of five important digital platforms supporting open access monographs. The content of participating platforms will be enriched with tools that enable identification, authentication and interoperability (via DOI, ORCID, Fundref), and tools that enrich information and entity extraction (INRIA (N)ERD), the ability to annotate monographs (Hypothes.is), and gather usage and alternative metric data. This paper focuses on the development and implementation of Open Source Metrics Services that enable the collection of OA Metrics and Altmetrics from third-party platforms, and how the architecture of these tools will allow implementation in any external platform, particularly in start-up Open Access publishers.

INDEX

Keywords: Metrics, Open Access, Digital Monographs

AUTHOR

JAVIER ARIAS

Open Book Publishers, United Kingdom
javi@openbookpublishers.com
(corresponding author)