Shared Infrastructure for Next- Generation Books: HIRMEOS

Brian Hole, Francesco De Virgilio, Chealsye Bowley

Shared Infrastructure for Next-Generation Books: HIRMEOS

Brian Hole, Francesco de Virgilio and Chealsye Bowley

1 A central goal of HIRMEOS is to better integrate open access books with the open science ecosystem, in particular helping to increase the visibility and value of work from the social sciences and humanities (HIRMEOS 2018). Its wide-ranging scope includes building prototypes and ongoing systems to provide full integration with authentication and interoperability services (DOIs, ORCiD and FundRef), content enrichment services (INRIA N(E)RD), dissemination platforms (DOAB), a new peer review certification platform, annotation services (hypothes.is), and usage and alternative metrics.

2 A key aspect of this project is its focus on building fully open source, shared infrastructure. This is not simply due to altruistic community principles (though these are implicit), nor a matter of checking a box to qualify for funding, but also a logical and critical commercial decision by the nonprofit and for-profit project partners involved. There is a cultural change underway within the open access movement for shared and open infrastructure. The Collaborative Knowledge Foundation's PubSweet project (Coko Foundation 2018) for example sees commercial open access publishers such as Ubiquity Press, ELife and Hindawi working together to produce the next generation of fully open source journal publishing software, using best practices in collaborative open source development (see Hyde 2017). The rationale is that the publishers involved share the cost (and risk) of developing core systems that they all require, and are able to benefit from more cost-effective, up-to-date and tailored software than they would otherwise have had access to through traditional suppliers. Risk is further lowered in that key systems remain shared rather than being potentially acquired by larger market players that may not be equally incentivized to retain certain aspects, such as open source code or open access functionality. These concerns are shared more broadly by the library community as well, as articulated recently for example in the proposal by David Lewis of the University of Indiana (Lewis 2017) that libraries contribute 2.5% of their budgets towards shared open infrastructure.

3 HIRMEOS aims to play this role with shared infrastructure for book publication, equally benefiting for-profit publishers such as Ubiquity Press alongside non-profit ones such as EKT Open Book Press, OpenEdition Books, Open Book Publishers and Göttingen University Press and OApen. The following will focus on the Ubiquity Press led annotation and altmetrics deliverables of the project as examples.

4 User-generated annotation (as opposed to machine-generated enhancement) has been described as a "scholarly primitive" activity (Unsworth 2000) embedded throughout the scholarly lifecycle. As consumption of books moves online it is therefore essential to enable such fundamental activity if users are to adopt new platforms. At the same time such activity is naturally useful to measure if we are interested in the degree to which a book is being truly used within that lifecycle.

5 Previous attempts to integrate annotation with online systems were not entirely successful, largely due to limitations of browsers and the web (Bradley 2007). This has since changed with the introduction of W3C standards for web annotation (W3C 2017) enabling a range of new annotation services such as hypothes.is and PaperHive to be reliably used across the web.

6 The annotation work in HIRMEOS involves ensuring that the hypothes.is system is integrated with all of the project partners' products, providing for a consistent user experience across all platforms. Among other technical challenges, this involves ensuring that hypothes.is is able to work with the EPUB readers used by these platforms, as this often involves iteratively embedded objects that are difficult for such software to access. Metrics for annotations made will be collected and stored by the metrics system also being developed as part of the project.

7 Altmetrics, which aim to look beyond traditional measures such as peer review, citation counts and Impact Factors (Priem *et al* 2010) have become increasingly salient within STEM publication within recent years (Warren 2017, Piwowar 2013). The fact that they are now increasingly used in academic assessment programs leads many to worry that HSS scholars are at a disadvantage in regard to how their work is measured and thereby valued (e.g. Hammarfelt 2010). Altmetrics are therefore an important service to be provided by HIRMEOS, with the aim of enabling the use and impact of open access books to be assessed in a comparable way to works from other disciplines.

8 Through the project a metrics service based on that already used by Ubiquity Press is currently being developed (HIRMEOS 2017) and will be integrated with all partner platforms in order again to provide a familiar, consistent and comparable experience for users. The system makes use of DOIs to produce metrics for citations, as well as altmetrics from Twitter, Facebook, Wikipedia and hypothes.is. The publishers individually register their DOIs with the service hosted on Ubiquity Press servers (initially, just loading a CSV). The service then harvests metrics for those DOIs on a daily basis and these can be queried through an open API or displayed on each platform in a fully consistent way in time series, through a JavaScript widget embedded on their respective websites. Wherever possible the system makes use of existing open infrastructure (e.g. Crossref Event Data), and community-expected standards (e.g. COUNTER compliance).

9 Both services are being developed using the Django web framework and the Python programming language (respectively versions 2 and 3.6).

10 The altmetrics service consists of a shared set of functions (dealing with data aggregation, notifications, the API and some database-related tools) which provides a solid ground to a

flexible set of plugins, each one targeting a specific metrics source. Plugins are intended to be relatively small in size and are not tied to any specific database structure: their scope is limited to talking to a specific data source, gathering the relevant data and handing the results to the altmetrics service after reshaping them consistently. This allows any contributor to add a new plugin in order to integrate a new source. A generic demo plugin is available in the code base to make it easier for developers to create new plugins. Altmetrics related to annotations are harvested by a dedicated annotation plugin, which is itself part of the altmetrics service.

11 One of the challenges of the project is the wide variety of sources to be aggregated in order to gather significant statistics. In order to make the retrieval easier and more structured, all metrics are built around the DOI as the unique identifier of the document. This involves resolving the different identifiers used by different services to the same DOI. In order to do so, the altmetrics service relies on the registration of the DOIs by the HIRMEOS partners; the annotation plugin relies on the Crossref Event Data service (which in turn integrates well with hypothes.is).

12 Regarding the annotation plugin specifically, partners will be asked to implement standard Dublin Core meta tags in their pages; any annotation created by the reader on a page will be registered by hypothes.is using the URL of the page, and the Crossref Event Data service will take care of associating the two, making annotations available on a per-DOI basis.

13 The most complex aspect the altmetrics service is dealing with the large amount of DOIs it will be required to collect metrics for. A scalable approach has therefore been adopted, turning any data gathering process into an asynchronous task that is handed to a queuing system (RabbitMQ). The queuing service distributes tasks to all the available workers, which perform the actions of querying the data source and collecting the results. Each task is worked on independently from the others, and the number of workers connected to the queue can be automatically scaled based on the number of DOIs registered in the service, thus avoiding any performance degradation.

14 The altmetrics service has been designed with data portability in mind and aims to provide a standard way to add and extract data to and from the project. In this context, the REST API is one of the key features of the service, enabling programmatic access to the metrics. Using well known strategies for API caching (e.g. last-modified HTTP header), the API provides fast access to the metrics. The API will be primarily used by the web widget integrated in the partners' websites to show metrics, but can also be used for producing visualizations or other statistical purposes by the partners.

15 Currently, the only way provided to register new DOIs into the service is a manual CSV upload, but there are plans to extend the above-mentioned API interface to include DOI registration, allowing partners' publishing platforms to POST to the altmetrics service every time a new DOI is registered.

16 Given the number of different components the service is built on, and in the spirit of lowering the entry barrier for new developers, a Docker-based deployment approach has been chosen to run the service in production environments. Configuration files to build the Docker container are provided as part of the open source codebase, enabling any developer to have a fully working and configured altmetrics service in minutes. Other configuration files provided with the project (Compose stackfile) allow developers to

create a fully functional environment including all the back-end services needed to run the project (RabbitMQ, Redis cache, Celery workers).

17 Some practical limitations of the project are related to the frequency with which data sources are to be queried and the possibility that these could be limited by throttling. This scenario could become likely with the growth in the number of DOIs registered in the service. Currently, a configuration internal to the altmetrics service is used to schedule the queries, which are performed daily for all DOIs. As the service is still under active development, tests will be performed to decide what the most optimal frequency is. Some of the measures which could be taken to mitigate the throttling issue are to query data sources for multiple DOIs in a single request (where possible), or use remote agents running on separate hardware instances (having different IPs), distributing the data gathering workers in order to avoid queries from the same IP (the queuing strategy makes this approach very easy). As Facebook discontinued API access to public content (Facebook 2018); overcoming this issue will require an investigation of the different possibilities, which has not been conducted yet.

18 Backend and integrations tests are being developed as part of the project. Testing is a standard practice in the software development process, but it becomes even more important when the service relies on integrations with external data providers. At completion, each data source plugin will include test fetching data from the data provider, ensuring that the code is still able to deal with any change in the data provider API. The project is currently hosted on HIRMEOS GitHub's repository, and tests are run automatically on the Ubiquity Press internal continuous integration service before release of any new features, which ensures that any change in external providers is addressed before the release.

19 The addition of annotations as an altmetric is a key highlight of the project. As mentioned above annotation is a fundamental activity within scholarship, yet it is not well measured or reported. It can be argued that it is more widespread and deeply practiced within HSS disciplines in particular, as books often require a longer and deeper engagement from researchers than do research articles, necessitating more in depth note-taking and receiving less immediate citation. It is hoped that by providing an altmetric to measure this activity, the metrics of books will be boosted in the longer term, enabling them to be better compared to STEM outputs that to date have been more easily and immediately assessed. A potential future expansion of the service may also enable the system to achieve one of the original goals of altmetrics as envisioned by Priem et al (2010), i.e. to "look beyond counting and emphasize semantic content", based on the location, frequency and content of the annotations.

BIBLIOGRAPHY

References

Coko Foundation. (2018). *PubSweet website*. https://pubsweet.org/. Accessed 9 January 2018.

Facebook. (2018). *Public Feed API*. Available at: https://developers.facebook.com/docs/public_feed

Hammarfelt, B. (2010). "Interdisciplinarity and the Intellectual Base of Literature Studies: Citation Analysis of Highly Cited Monographs." *Scientometrics*, 86: 705–725. http://doi.org/10.1007/s11192-010-0314-5.

HIRMEOS. (2017). *Deliverable D6.1: Metric Services Specification.* Available at: http://www.hirmeos.eu/wp-content/uploads/2017/11/HI61-Metrics_Service_technical_specification-final.pdf. Accessed 9 January 2018.

HIRMEOS. (2018). *HIRMEOS Project website*. http://www.hirmeos.eu/about-hirmeos/ Accessed 9 January 2018.

Hyde, A. (2017). *The Cabbage Tree Method: Open Source Collaborative Product Development*. Version 0.2. Available at: https://www.cabbagetree.org/ Accessed 9 January 2018.

Lewis, D.W. (2017). *The 2.5% Commitment*. http://hdl.handle.net/1805/14063. Accessed 9 January 2018.

Piwowar, H. (2013). "Value All Research Products". *Nature*, 493: 159. https://doi.org/10.1038/493159a

Priem, J., Taraborelli, D., Growth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. http://altmetrics.org/manifesto. Accessed 9 January 2018.

Unsworth, J. (2000). *Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? Symposium on "Humanities Computing: formal methods, experimental practice,"* King's College, London, May 13, 2000. http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html. Accessed 9 January 2018.

W3C. (2017). *Three Recommendations to Enable Annotations on the Web*. https://www.w3.org/blog/news/archives/6156. Accessed 9 January 2018.

Warren H.R., Raison, N., Dasgupta, P. (2017). "The Rise of Altmetrics." *JAMA*. 317(2): 131–132. http://doi.org/10.1001/jama.2016.18346.

ABSTRACTS

This paper presents an introduction and status report on work being done to provide shared infrastructure for open access book publishers under the HIRMEOS (High Integration of Research Monographs in the European Open Science infrastructure) project. It focuses specifically on the

work being done to provide shared altmetrics services, including reporting on annotation activity.

INDEX

Keywords: monographs, metrics, altmetrics, annotation, open source software

AUTHORS

BRIAN HOLE

Ubiquity Press, United States
brian.hole@ubiquitypress.com
(corresponding author)

FRANCESCO DE VIRGILIO

Ubiquity Press, United Kingdom
francesco.devirgilio@ubiquitypress.com

CHEALSYE BOWLEY

Ubiquity Press, United States
chealsye.bowley@ubiquitypress.com