Scaling Usage Statistics across Repositories as an OpenAIRE Analytics Service

Dimitris Pierrakos, ATHENA Research & Innovation Center, dpierrakos@gmail.com
Jochen Schirrwagen, Bielefeld University, jochen.schirrwagen@uni-bielefeld.de
Pedro Miguel Oliveira Bento Príncipe, University of Minho, pedroprincipe@sdum.uminho.pt
Ricardo Saraiva, University of Minho, rsaraiva@sdum.uminho.pt

Abstract

Usage metrics about scholarly output, such as publications and research data, are one of the measures to assess Open Access impact. The OpenAire 2020 [1] project aims to offer a service that monitors and analyzes usage information, as well as exploits usage metrics like views and downloads, which could be used as complements of bibliometrics and webometrics. In this paper, we present the first step towards the implementation of this service, manifested as a pilot run in a set of repositories, together with some initial results which illustrate the use of the applied methodology.

Keywords
Open Access, repositories, usage analytics, Piwik

Introduction

Tracking the usage of their resources and providing usage statistics is one of the value added services of Open Access repositories. Usage tracking involves a number of parameters, such as the requested item, the timestamp of the request, the id of the visitor, etc.

OpenAIRE2020 aims to facilitate the above task by monitoring and analyzing usage data of the repositories and exploiting a number of usage metrics like downloads and metadata views. The results of this process will be finally used to examine correlations with other metrics, e.g. bibliometrics and webometrics. Moreover, and being aware of the sensitivity of the usage data, the legal constraints will be considered with regard to the EU data privacy policy and national regulations. The final outcome will be an OpenAIRE service for tracking, collection, cleaning, analysis, evaluation and COUNTER-compliant reporting of the usage data. This service will be provided both in the OpenAIRE portal and via an API. In this paper we present a short description of the Usage Statistics service, together with its first exploitation via a pilot on Open Access repositories, as well as a set of initial results.

Usage Analysis in OpenAire 2020

OpenAIRE2020 represents a pivotal phase in the long-term effort to implement and strengthen the impact of the Open Access (OA) mandate of the European Commission (EC). Such impact can be measured by analyzing the usage information of the Open Access repositories. OpenAIRE2020 though, not only collects and process such information, but also extends and scales such service on a European and ultimately at an international level.

OpenAIRE2020 tackles another important aspect, which is the duplication of documents across the repositories. In particular, the aggregation of bibliographic metadata records across repositories leads to what is known as duplicates, i.e. groups of records referring to the same document. OpenAIRE2020 offers the *deduplication* of such records [1], by scanning the whole collection to identify such groups and eventually replace them with only one record (i.e. "merging"), which unambiguously describes the publication in the collection. In OpenAIRE2020 usage information is collected using the open source Piwik platform [3]. A short description of the methodology is presented in the following section.

Usage Analysis Methodology

Two approaches are foreseen for the collection of usage data, named Tier 1 and Tier 2. Tier 1, depicted in Figure 1, is the default workflow offered by the Usage Analytics service in OpenAIRE2020. Tier 1 exploits a workflow provided by Piwik platform. Open Access repositories add JavaScript tracking page tags which will notify OpenAIRE's gathering service of the usage events. Anonymization services will also be offered by OpenAIRE, using the Piwik anonymization protocol, whenever it receives new usage data and before it executes any further processes on them.

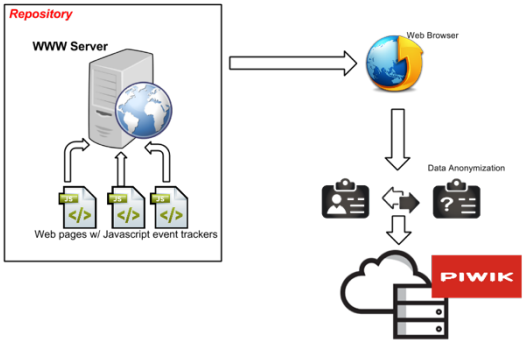

Figure 1. Usage event tracking from repositories

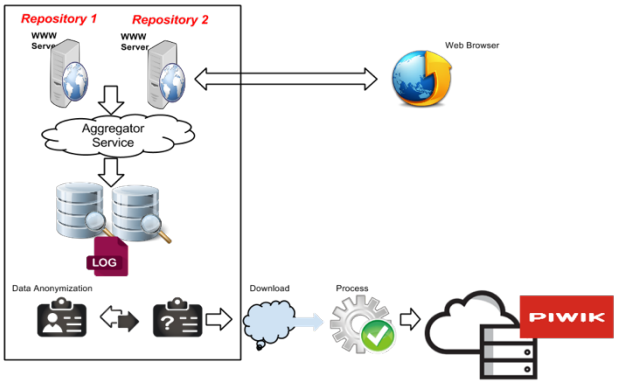

Figure 2. Collecting consolidated usage statistics

A different approach for the usage analytics service, named Tier 2 is depicted in Figure 2, whereas data providers or usage statistics aggregation services (e.g. IRUS-UK [4]) offer a bulk download method for the usage data. Moreover, Tier 2 approach supports the gathering of consolidated statistics reports using other protocols such as SUSHI-Lite [5].

Usage Statistics pilot with Portuguese repositories

As mentioned in the introductory section, the first step towards the implementation of the usage analytics service is a pilot run in three Portuguese repositories, following the Tier 1 approach. The repositories participating in the pilot comprises:

- The institutional repository of Minho University (http://repositorium.sdum.uminho.pt) with 26739 documents in OpenAIRE and tracked from 1/7/2015 to 31/12/2015.
- The Estudo Geral, the repository of the University of Coimbra (https://estudogeral.sib.uc.pt) with 13043 documents in OpenAIRE and tracked from 21/9/2015 to 31/12/2015.
- The repository of University of Évora (http://dspace.uevora.pt/rdpc) with 8230 documents in OpenAIRE and tracked from 21/9/2015 to 31/12/2015.

In the following figures we show the initial results of the pilot phase. Figure 3 presents the metadata views and downloads of the articles of the tracked repositories. In Figure 4, we present the accumulated information of same publications but tracked from two different repositories.

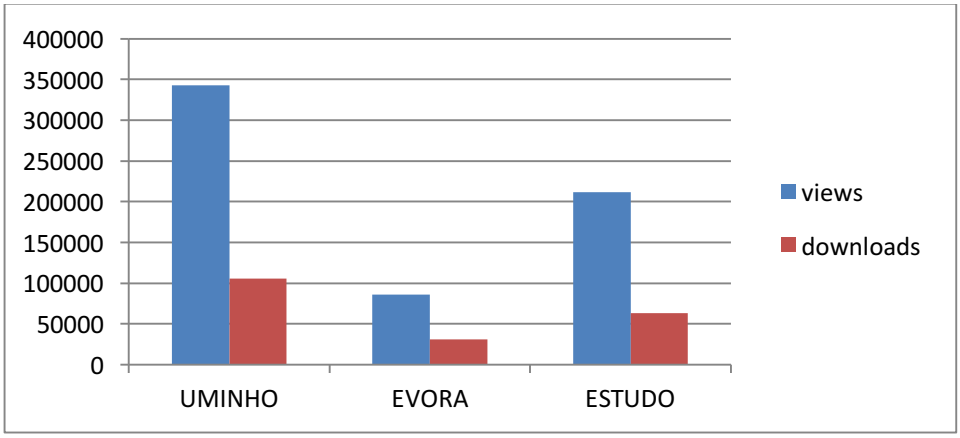

Figure 3 Metadata Views – Downloads on Pilot Repositories

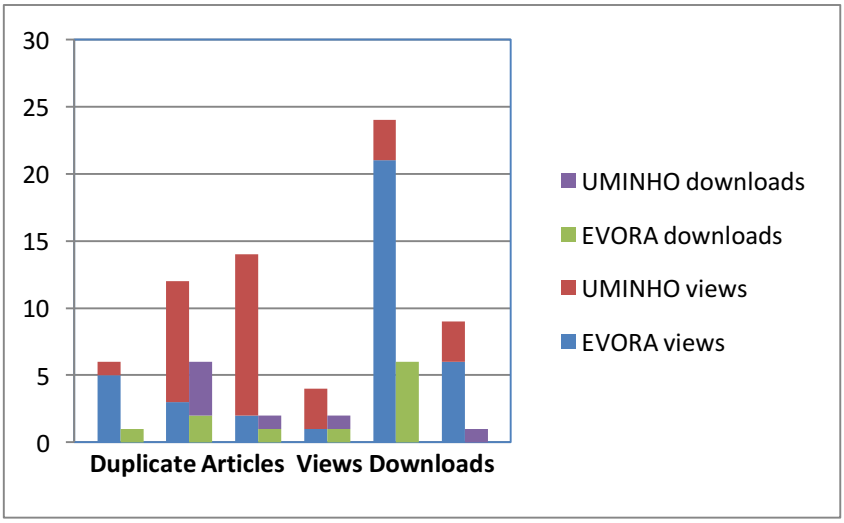

Figure 4 Metadata Views – Downloads of duplicate information

Conclusion

In this paper we presented the initial steps taken towards the implementation of the usage analytics service in OpenAIRE2020. We described the methodology and the results of the pilot phase that we implemented. Next steps include approaching other types of data providers (OA journal platforms, CRIS), offering the statistics service as a data source for other services such as the Lagotto application [6].

References

[1] www.openaire.eu

[2] Manghi, Paolo, Marko Mikulicic, and C. Atzori. "De–duplication of aggregation authority files." International Journal of Metadata, Semantics and Ontologies7.2 (2012)

[3] http://piwik.org/

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

[5] http://www.niso.org/workrooms/sushi/sushi_lite/

[6] http://www.lagotto.io/