# D8.5 USAGE STATISTICS SERVICES


12/2016 | OpenAIRE2020

Open Access Infrastructure for Research in Europe towards 2020
Deliverable Code: D8.5 - Version 1
PUBLIC

This deliverable specifies the Usage Statistics Service.


H2020-EINFRA-2014-1
Topic: e-Infrastructure for Open Access
Research & Innovation action
Grant Agreement 643410

# Document Description

## D8.5 – Usage Statistics Services

| | |
|---|---|
| WP8 - Information Space maintenance and extension | |
| WP participating organizations: CNR, ICM, CERN, ARC, UNIBI, UMINHO, UBONN, JISC, DANS-KNAW, COUPERIN, EKT | |
| Contractual Delivery Date: 12/2016 | Actual Delivery Date: 04/2017 |
| Nature: Report | Version: 1 Final |
| Public Deliverable | |

### Preparation Slip

| | Name | Organisation | Date |
|---|---|---|---|
| **From** | Jochen Schirrwagen | UNIBI | 9/12/2016 |
| | Dimitris Pierrakos | ARC | |
| | Ross MacIntyre | Jisc | |
| | Paul Needham | Jisc/IRUS-UK | |
| | Georgi Simeonov | IMI-BAS | |
| | Pedro Principe | UMINHO | |
| | André Dazy | Couperin | |
| **Edited by** | Jochen Schirrwagen | UNIBI | |
| **Reviewed by** | Aenne Loehden | UNIBI | 05/04/2017 |
| | Tony Ross Hellauer | UGOE | 03/04/2017 |
| | Daniel Beucke | UGOE | 04/04/2017 |
| | Joseph Green | UCD | 05/04/2017 |
| **Approved by** | | | |
| **For delivery** | Mike Chatzopoulos | UoA | |

### Document Change Record

| Issue | Item | Reason for Change | Author | Organization |
|---|---|---|---|---|
| V0.9.2 | First Draft | Outline and initial specifications | Jochen Schirrwagen, Dimitris Pierrakos | UNIBI ARC |
| V0.9.9 | Second Draft | Incorporated comments; updated with guidelines | Jochen Schirrwagen | UNIBI |

| | | | |
|---|---|---|---|
| V1.0 | Final version | Jochen Schirrwagen | UNIBI |

# Table of Contents


**1| INTRODUCTION 8**

**1.1 CONSIDERATIONS REGARDING DATA PROTECTION AND PERSONALLY IDENTIFIABLE INFORMATION 8**
1.1.1 GENERAL DATA PROTECTION REGULATION (GDPR) OF THE EUROPEAN UNION 12
**1.2 OVERVIEW OF EXISTING USAGE STATISTICS SERVICES, TOOLS AND MEANS OF USAGE EVENT COLLECTION 13**
**1.3 STANDARDS FOR USAGE STATISTICS - COUNTER REPORTS AND SUSHI-LITE 16**

**2| USAGE ANALYTICS SERVICE SPECIFICATION 18**

**2.1 REGISTRATION OF THE REPOSITORY IN THE OPENAIRE PIWIK PLATFORM 19**
**2.2 LOGGING OF USAGE EVENTS IN THE PIWIK PLATFORM 20**
**2.3 APPLYING COUNTER RULES ON USAGE EVENTS 22**
**2.4 STORAGE IN THE OPENAIRE STATISTICS DATABASE 24**
**2.5 REPRESENTATION AND PROVISION OF USAGE STATISTICS 25**
**2.6 PIWIK REPORTING SCHEME 29**
**2.7 SCALABILITY OF PIWIK 30**
**2.8 GATHERING USAGE REPORTS FROM DATA PROVIDERS SUPPORTING THE SUSHI-LITE PROTOCOL 30**

**3| RESULTS FROM PILOT IMPLEMENTATION 33**

**3.1 COMPARISON OF USAGE ACTIVITY BETWEEN PIWIK AND GOOGLE ANALYTICS 35**

**4| DISCUSSION AND NEXT STEPS 39**

**5| REFERENCES 40**

**6| APPENDIX 41**

**6.1 PARTICIPATING REPOSITORIES IN OPENAIRE'S PIWIK 41**
**6.2 LIST OF BOTS 42**
**6.3 OPENAIRE GUIDELINES FOR COLLECTING USAGE EVENTS AND PROVISION OF USAGE STATISTICS V1 43**
6.3.1 PURPOSE 43
6.3.2 SCOPE OF APPLICATION 43
6.3.3 USAGE DATA COLLECTION, PROCESSING AND REPORTING 43
6.3.4 PARTICIPATION AND WORKFLOW 44
6.3.5 RESPONSIBILITIES 45
6.3.6 SOFTWARE SUPPORT 45
6.3.7 MAINTENANCE OF THE GUIDELINES 45


# Table of Figures


*Figure 1: Different options of Tracking or logging of usage events in a repository 15*
*Figure 2: Usage statistics service architecture components 19*
*Figure 5 OpenAIRE statistics DB 25*
*Figure 6: usage per publication 26*
*Figure 3: real time usage activity 29*
*Figure 4: real time visitor map 30*
*Figure 7: number of views and downloads of tracked publication on repository level 33*

*Figure 8: accumulated number of views and downloads of tracked publications hosted in different repositories* ____ *34*
*Figure 9: views and downloads in the OpenAIRE portal and tracked repositories participating in the pilot for the period 01-jan-2016 – 10-dec-2016* ____ *35*
*Figure 10: Usage activity in piwik* ____ *37*
*Figure 11: usage activity in google analytics* ____ *38*

# Disclaimer

This document contains description of the OpenAIRE2020 project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the OpenAIRE2020 consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)


OpenAIRE2020 is a project funded by the European Union (Grant Agreement No 643410).

# Acronyms

| | |
|---|---|
| COUNTER | COUNTER (Counting Online User NeTworked Electronic Resources) is an organization which develops standardized methods and reports for measuring the use of electronic resources. |
| COUNTER Code of Practice | A COUNTER compliant standard for the recording, consolidation and reporting of usage at the individual article level |
| NISO | National Information Standards Organization |
| OJS | Open Journal Systems |
| SUSHI | Standardized Usage Statistics Harvesting Initiative |

# Publishable Summary

This deliverable report on “Usage Statistics Services” is carried out under the workpackage "Information Space maintenance and extension".

It aims to align with existing guidelines and standards for the collection, cleaning and processing of usage data from OpenAIRE data providers, such as literature and data repositories, e-journal platforms and CRIS, and eventually for the provision of consolidated usage statistics in the OpenAIRE portal, repository dashboard and via an API for external services.

The report documents the current and upcoming policies on data protection, which must be considered when collecting and evaluating usage statistics. They are compared on the level of the EU and of selected national policies.

Those standards, guidelines and policies are then considered in the specification of the “Usage Statistics Service”, which aims to seamlessly integrate in the existing OpenAIRE infrastructure.

As a proof of concept several pilots with repositories from Portugal and other countries, e-journal platforms and the IRUS-UK service have been run. While those pilots have revealed several issues in the tracking code they have also informed and improved the implementation of the tracker code for repositories and the development of the OpenAIRE Sushi-Lite API for the provision of usage statistics.

# 1 | INTRODUCTION

A usage statistics service, which allows the sharing of statistics across repositories, e-journals and CRIS, will provide significant added value for different stakeholders.

On the data-provider level, it can serve repository managers and hosting institutions as a tool to evaluate the success of the publication platform.

On the individual item level, it can demonstrate popular publications to authors and readers. In addition to other traditional (e.g. citation counts) and alternative metrics (e.g. mentions, recommendations) it can inform funding authorities in research evaluation processes.

Usage statistics on the item level can reflect relevance of a particular research output, of topics, of (disciplinary) data sources over the course of time and up to the present, e.g. they are an important indicator to analyze trends.

For non-traditional output types (e.g. research data, research software), usage statistics are often the only indicator available, while the implementation of data citation standards lags behind.

The OpenAIRE Usage Statistics Service aims to facilitate the above added-value services by tracking, collecting, analyzing and monitoring usage data from its network of data providers and exploiting usage metrics like downloads and metadata views. Moreover, OpenAIRE's distributed network of data providers allows the aggregation of usage data about publications which are published in several places.

Being aware of the sensitivity of this usage data, legal constraints will be considered regarding the EU Data Protection Directive and policies on the national level.

The aim is to allow COUNTER-conformant reports on usage statistics to be generated and thus enable the results of this process to be used to examine correlations with other types of metrics, e.g. bibliometric and webometric. This service will be integrated with the repository dashboard, the OpenAIRE portal and API for 3rd party reuse.

## 1.1 Considerations regarding Data Protection and Personally Identifiable Information

Usage statistics are based on the processing of usage events. Usage events are triggered by accessing digital objects in repositories, e-journals and CRISs. A triggering client can be a machine (e.g. search bot) or a human. A usage event comprises information about the accessed object (e.g. URL, URN, DOI, OAI-record identifier), timestamp of the event, status of the access but also information about the client triggering the event, including IP address and other client information transferred as part of the request (e.g., "user-agent"). Data protection policies take effect in case of personally identifiable or sensitive personal information.
The interpretation of e.g. the IP address (and in combination with additional attributes and processing stages) as personally identifiable information in the EU and EU member states varies. This section reports and compares the current legal status by example of selected EU member states and on the level of the EU itself. This is important insofar as for repositories national regulations apply and usage events are transferred to the OpenAIRE Usage Statistics Service as a 3rd party operated in another country. Therefore, the configuration of the OpenAIRE usage statistics service will be described in order to comply with the legal conception.

OpenAIRE is an infrastructure operated in the European Union and as such it must adhere to the requirements of the EU Data Protection Directive[1]. This directive will be replaced by the General Data Protection Regulation in 2018.

In the following selected policies on the national level are explained in more detail and concluded by the policy of the EU.

Germany

In Germany the processing of personal data is in principle governed by the general data protection laws (the Federal Data Protection Act[2] as well as the data protection laws of the German federal states), unless special regulations apply. The Federal Data Protection Act implements the EU Data Protection Directive http://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:31995L0046 .

For the OA-Statistik project a legal opinion[3] was prepared in 2011 which states that IP addresses can be personally identifiable information and therefore a statistics service implementation must process hashed salted IP addresses.

Bulgaria

In Bulgaria the "Law for Protection of Personal Data" has been in force since 01.01.2002. The official document (translated in English) with full text of the law is published on the website of the Bulgarian Commission for Personal Data Protection (CPDP) at:
https://www.cpdp.bg/en/index.php?p=element&aid=373
Article 2 (suppl., - SG 70/04, in force as from 01.01.2005; amend. - SG 103/05)

(1) (amend. - SG 91/06) "Personal data" shall refer to any information relating to an individual who is identified or identifiable, directly or indirectly, by reference to an identification number or to one or more specific features.

(2) Personal data must be:

1. processed in legal compliance and in a bona fide manner;
2. (suppl. - SG 81/11) collected for specific, precisely defined and legal purposes and not be submitted to additional processing in a manner incompatible with such purposes; additional personal data processing for historical, statistical or research purposes shall be allowed provided the data controller has ensured proper protection, guaranteeing that the data are not being processed for any other purposes, except the cases explicitly provided for in this Law;
3. (amend. - SG 91/06) proportionate to the purposes for which they are being processed and not exceeding their scope;
4. accurate, and updated if necessary;
5. deleted or corrected when found to be imprecise or disproportionate to the purposes for which they are being processed;

[1] EU Data Protection Directive: http://ec.europa.eu/justice/data-protection/
[2] https://www.gesetze-im-internet.de/englisch_bdsg/
[3] Zendas legal opinion: https://dini.de/fileadmin/oa-statistik/gutachten/ZENDAS_Gutachten_2011.pdf

6. maintained in a form that enables identification of the respective individuals for a period not exceeding the time necessary for the purposes for which such data are being processed; personal data which will be stored for a longer period of time for historical, statistical or research purposes shall be kept in a format precluding the identification of individuals.

The law does not explicitly designate internet protocol addresses (IP) as personal data. IPs can have different statuses depending on whether they can identify individuals. In one case an IP can be personal data when it's possible to track and identify the individual. In another case IPs are not personal data as it is not possible to identify the person behind a given address or number. For example, the same public IP address can be shared with many people, using NAT (Network address translation), Proxy servers, VPN etc.
The service owner who collects, uses, processes and stores personal data must be registered in the country as an administrator of personal data and follow rules, restrictions and all regulations from the "Law for Protection of Personal Data".

Great Britain
An IP address in isolation is not personal data under the UK Data Protection Act[4], according to the Information Commissioner. But an IP address can become personal data when combined with other information or when used to build a profile of an individual, even if that individual's name is unknown.

The IRUS-UK[5] service uses IP addresses solely to help to identify robotic and rogue usage. IRUS-UK exposes and shares COUNTER statistics, not raw download data, so those IP addresses are never used, exposed or shared for any other purpose than detection of robotic or fraudulent events.
As far as the service is able to determine the usage is reasonable and compliant with the recent Court of Justice of the European Union ruling that personal IPs can't be stored, *unless to thwart cybernetic attacks or similar*[6].

Portugal
The law governing the processing of personal data in the context of publicly available electronic communications networks and services is Law 41/2004 of 18 August, which implemented Directive 2002/58/EC on the protection of privacy in the electronic communications sector, as amended by Law no. 46/2012, transposing Directive no. 2009/136/EC of the European Parliament and Council, of 25 November. The Law 32/2008, of 17 June, which implemented

[4] http://www.legislation.gov.uk/ukpga/1998/29
[5] IRUS-UK: Institutional Repository Usage Statistics UK
[6] http://curia.europa.eu/juris/document/document.jsf?text=&docid=184668&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=1075747

Directive 2006/24/EC of the European Parliament and Council of 15 March 2006, is on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or public communications networks. Portuguese law sets out a very broad concept of data processing: any operation or use involving personal data is considered to be data processing. This includes the collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, deletion or destruction of personal data. The legal terms for the use of cookies and their data privacy implications have been set out in the laws concerning privacy in the electronic communications sector. The current rules determine that the use of cookies requires express and informed consent (opt-in) from data subjects (an opt-out option is not sufficient). Therefore, data controllers must obtain consent and provide clear, adequate and complete information regarding the purpose of their use of cookies. The only exception to this general opt-in rule is when the storage or access is technically and strictly necessary for the legitimate purpose of allowing access to a service that has been specifically and expressly requested by the data subject (that is, the subscriber or user). Data transfer to other countries within the EU does not require prior authorisation from the CNPD (Portuguese Data Protection Authority). In fact, in the set of laws internet addresses or IP are not explicitly mentioned as personal data.

France

The law, Act n°78-17 of 6 January 1978 on information technology, data files and civil liberties (translated in English behind the link below), relates to data protection in France:

https://www.cnil.fr/sites/default/files/typo/document/Act78-17VA.pdf

When processing personal data, the French Data Protection Authority (CNIL: https://www.cnil.fr/professionnel) must be informed and asked for authorisation via a CIL (civil liberties correspondent). This also applies for IP addresses which are interpreted as personal data.

Greece

The existing data protection regime is as follows: Law 2472/1997 (Data Protection Law), as amended by Law 3471/2006, incorporated in the Greek legislation of the EU Data Protection Directive (officially Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data). The Data Protection Law grants the data subjects, i.e. individuals (any natural person to whom such data refer and whose identity is known or may be found based on specific factors) certain rights and imposes certain responsibilities on data controllers, i.e. anyone who keeps personal data in a file and processes them.

For the purposes of the Data Protection Law, "personal data" shall mean any information relating to the data subject.

Consolidated statistical data are not considered personal data, when data subjects may no longer be identified.

The Data Protection Law established the Hellenic Data Protection Authority (HDPA), a constitutionally consolidated independent authority, responsible for the implementation of the data protection legislation (Laws 2472/1997 and 3471/2006), in particular the protection of individuals from the unlawful processing of their personal data.

Conclusion

The national policies on data protection and their implementation differs largely. From a legal point of view this is a potential problem not only for a cross-country usage statistics aggregation service but also for the data providers operating under their respective national law. However it is foreseeable that this situation will be more harmonized once the EU General Data Protection Regulation takes into force in 2018. Usage data contain personal identifiable information (by help and in combination of IP addresses, session identifiers, user agent information, data and time of requests). It is therefore strongly recommended to establish a culture of data sensitivity, use and configure tools that comply with data privacy policies at large, and that store and evaluate only information absolutely necessary while making use of appropriate anonymization of personal identifiable information.

### 1.1.1 General Data Protection Regulation (GDPR) of the European Union

The General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) is a regulation by which the European Commission intends to strengthen and unify data protection for individuals within the European Union (EU). It also addresses export of personal data outside the EU. The Commission's primary objectives of the GDPR are to give citizens back the control of their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU. When the GDPR takes effect, it will replace the data protection directive (officially Directive 95/46/EC) from 1995, mentioned above. It is a new directive as well as a new regulation; it will apply to police procedures, which will continue to vary from one member state to the other. The regulation was adopted on 27 April 2016. It enters into application on the 25th of May 2018 after a two-year transition period and, unlike a directive, it does not require any enabling legislation to be passed by national governments. Its provisions will be directly applicable in all member states.

The regulation applies if the data controller, or processor (organization), or the data subject (person) is based in the EU. Furthermore (and unlike the current directive) the regulation also applies to organizations based outside the European Union if they process personal data of EU residents. The regulation does not apply to the processing of personal data for national security activities or law enforcement ("competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties").

For the purposes of the regulation, 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

According to the European Commission personal data is any information relating to an individual, whether it relates to his or her private, professional or public life. It can be anything from a name, a photo, an email address, bank details, posts on social networking websites, medical information, or an IP address.

Further reference to the General Data Protection Regulation can be made to the following like:

http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EL

## 1.2 Overview of Existing Usage Statistics Services, Tools and Means of Usage Event Collection

Usage statistics services are offered by a number of providers. A well-known provider is IRUS-UK[7], which is a national aggregation service for UK Institutional Repository Usage Statistics, funded by Jisc. IRUS-UK collects raw download data from UK institutional repositories for all item types within repositories, such as articles, technical reports, exam papers, etc. It subsequently processes the raw data into COUNTER-conformant statistics.

A usage statistics service on the institutional repository level has been developed by the University of Minho for its DSpace repository RepositóriUM. The system was developed by exploiting the principles and some of the components of the statistics application created and released by the Australian National University (ANU). Despite the fact that the development was made to respond to the specific needs of RepositóriUM, the system is completely adjustable to other environments.[8]

Bepress, formerly the Berkeley Electronic Press, offers Digital Commons[9], an institutional repository and publishing software suite that allows institutions to showcase and preserve their scholarly output. Digital Commons supports Google Analytics for usage activity tracking.

PLOS One uses Apache log files to obtain usage activity such as views and downloads, and exploits COUNTER rules for reporting. SpringerLink and Elsevier also offer COUNTER Compliant Statistics for their journals.

From current usage statistics services different mechanisms to gather usage data or to collect consolidated usage statistics reports can be identified:

1) Usage data is recorded in web server log files and provided as a dump. Statistics services can collect the dump by OAI-PMH harvesting or by transmission via FTP.
2) Another method is based on tracking usage events in real-time. In this case data providers embed a tracking code in their web pages (e.g. tracked by Google Analytics, Piwik) or install a tracking plugin (e.g. tracked by IRUS-UK).
3) The third method supports requesting usage statistics as predefined reports or filtered snippets via SOAP or REST web services with the Standardized Usage Statistics Harvesting Initiative (SUSHI) explained in section 1.3

[7] www.irus.mimas.ac.uk
[8] http://hdl.handle.net/1822/4803
[9] https://www.bepress.com/products/digital-commons/

Mechanism 1) is e.g. utilized by OA-Statistik[10]. This service operates a lasting infrastructure and addresses German repositories for the collection and processing of usage data. The transfer protocol is based on OAI-PMH. OA-Statistik provides an OAI dataprovider client software (https://github.com/gbv/OA-Statistik) that in principle can be used with any repository platform. Usage events are extracted from web server log files, encoded as XML OpenURL Context Objects and harvested by the OA-Statistik server. The harvesting is typically done on a daily basis, hence not in real-time. The main issue following this approach is that it does not scale well for large amounts of usage events. OA-Statistik follows COUNTER processing rules to calculate statistics for views and downloads. Repositories can access the OA-Statistik API which returns statistics for views and downloads in JSON format. The statistics can be embedded in the repository landing pages. The disclosure of the statistical information is regulated by a license[11] based on Creative Commons BY-NC-SA 3.0. [2]

Mechanism 2) is used by IRUS-UK[12] (Institutional Repository Usage Statistics). It is a live service in the UK with currently 124 participating repositories. It is designed to collect raw usage data from repositories which are processed into COUNTER-conformant[13] article level usage statistics. While only download requests are counted as usage events it allows for a very efficient transfer of log entries encoded as OpenURL querystring entries to a remote statistics server in push mode. Details are defined in the IRUS tracker protocol (http://irus.mimas.ac.uk/help/toolbox/TrackerProtocol-V3-2014-04-22.pdf ). Implementation of the tracker protocol is platform specific and available for e.g. EPrints (http://bazaar.eprints.org/392/ ) and DSpace. Support has also been implemented where possible in Figshare and Elsevier's research information system PURE. In order to reduce volumes of usage traffic, usage events initiated by robots, crawlers and spiders should be identified on the client side and not transmitted to IRUS-UK. The usage statistics generated by IRUS-UK are made available to participating repositories and other aggregators, eg PlumX, via the IRUS-portal and are exposed via the SUSHI-Lite API. [3, 4]

In France AnalogIST[14] is a service hosting the usage statistics software ezPAARSE[15], released under a GPL-compatible licence. It is able to mine, analyse and enrich the logs generated by reverse proxies (ezProxy, Biblio PAM, Squid, and Apache) which record access to academic and scientific publishers' platforms. As a web-based application, it has an online form and an API allowing both on-demand and automatized treatment of these logs.

[10] OA-Statistik, originally funded by the DFG, https://dini.de/projekte/oa-statistik/english/ , is operated by the gbv, (Common Library Network), https://www.gbv.de/Verbundzentrale/serviceangebote/oas-service/open-access-statistik-service

[11] License information (in German): http://www.dini.de/fileadmin/oa-statistik/projektergebnisse/OAS_Lizenz.pdf

[12] IRUS-UK, funded by Jisc, http://irus.mimas.ac.uk/

[13] IRUS-UK has not yet been formally audited by the COUNTER auditors so it cannot currently claim 'COUNTER-compliance'. However the service has undergone a preliminary review with the COUNTER auditors, which verified that the statistics are produced in according with COUNTER processing rules and are thus 'COUNTER-conformant'.

[14] AnalogIST: http://analogist.couperin.org/start_en

[15] ezPAARSE: https://github.com/ezpaarse-project/ezpaarse

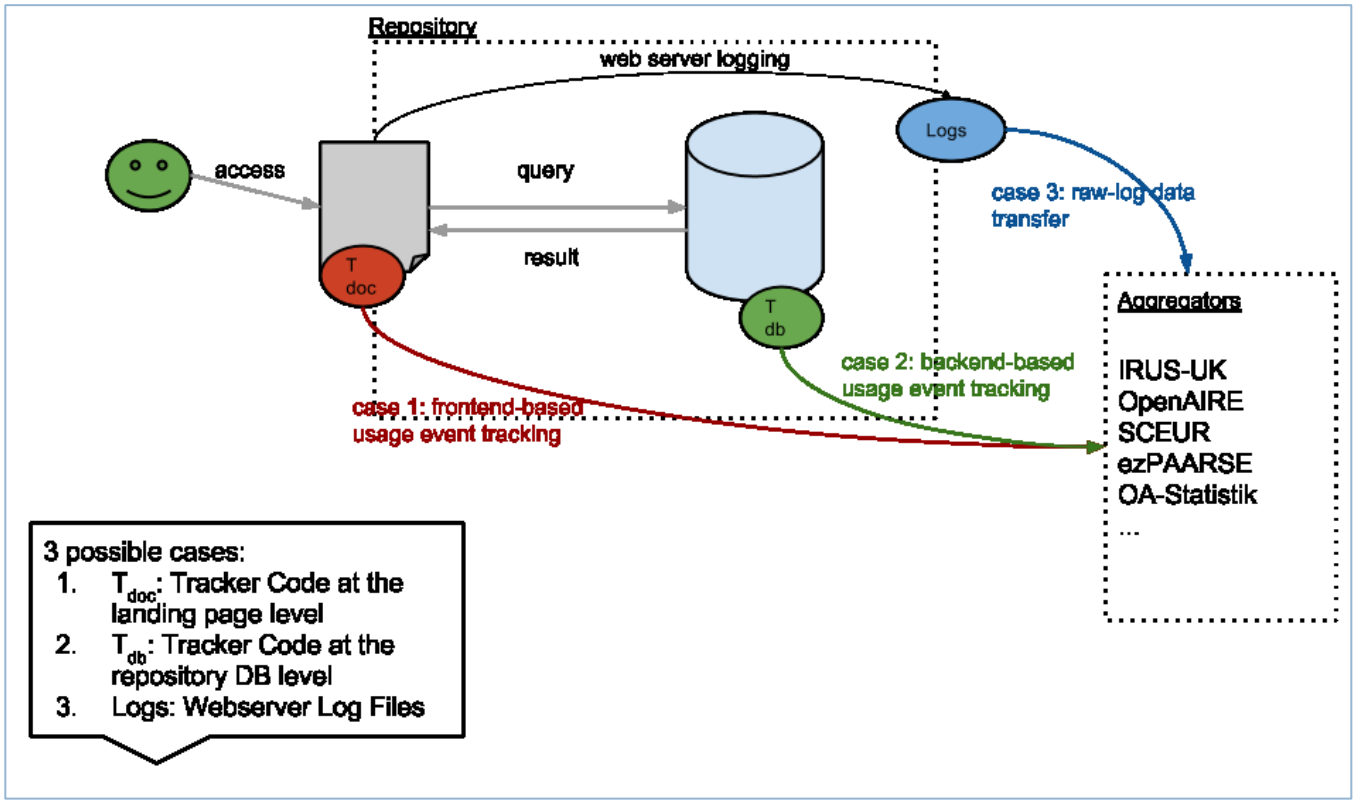


FIGURE 1: DIFFERENT OPTIONS OF TRACKING OR LOGGING OF USAGE EVENTS IN A REPOSITORY

Piwik is a platform for web traffic analysis which started in 2007 that has recently become the world's leading open-source analytics platform. Piwik provides users with valuable insights into their website traffic and visitors activity. The main advantage of Piwik against other analytics platforms, e.g. Google Analytics, is its approach to ownership of collected data. All information gathered, is available and controlled only by Piwik users and by default is not shared with any third parties.

Piwik supports both options – usage event tracking and import from webserver log files. Piwik is self-hosted, so the platform is stored on user's infrastructure and all data is tracked inside the user's database. Thus, the user keeps full data ownership and can control who has access. Piwik provides greater flexibility, as it allows the collection and storage of PII (personally identifiable information) and other sensitive data that cannot usually be stored outside of a user's system. Due to its privacy policy, Piwik is compliant with EU regulations, and is recommended by independent centers for Privacy Protection, e.g. ULD[16] in Germany and CNIL[17] in France.

Apart from its privacy policies, Piwik exceeds the main competitive Google Analytics platform in other respects, as shown in Table 1 below and described in [6]:

TABLE 1: COMPARISON OF PIWIK AND GOOGLE-ANALYTICS

| | Piwik | Google Analytics |
|---|---|---|
| Service Access | Open Source, self-hosting | Free to use; service provider solution |

[16] http://piwik.org/blog/2011/03/piwik-can-be-used-in-compliance-with-data-protection-laws/
[17] http://piwik.org/blog/2014/10/cnil-recommends-piwik-analytics-tool-no-cookie-consent/

| | | |
|---|---|---|
| Number of Hits per Month | Unlimited | 10 million |
| Number of user accounts per login | Unlimited | 10 |
| Data storage time | Unlimited | 25 months |
| Number of properties (websites, apps etc.) tracked per account | Unlimited | 50 |
| Custom Variables | 5 | 5 |
| Data Export | Unlimited | 5000 rows |
| Real time Analytics | Piwik offers real-time web analytics in all of its reports. | GA monitors user activity right after it happens, although period of delay is not explicitly stated. |

## 1.3 Standards for Usage Statistics - COUNTER Reports and SUSHI-Lite

COUNTER is an international organization to develop standards guidelines for the recording, reporting and exchange of usage statistics for electronic resources. COUNTER Codes of Practice address the terminology, the data processing, the format and layout of usage reports, the exchange protocol, auditing of vendors, compliance and governance. They aim for consistent, credible and compatible vendor-generated usage statistics. The current Code of Practice is Release 4[18] [5] while Release 5 is in preparation and will go into effect in January 2019[19].

COUNTER Usage Reports are defined on a monthly basis for different item data types, such as journals, databases, books, multimedia content, full-text items and for different types of metrics, such as full-text item requests, result clicks and record views.[20]

IRUS-UK has contributed to the application of COUNTER principles to article-level statistics thus enabling standardized usage statistics for institutional and thematic repositories[21], which will be incorporated into the COUNTER Code of Practice Release 5.

The Standardized Usage Statistics Harvesting Initiative (SUSHI) protocol has been developed by NISO and incorporated into the COUNTER Code of Practice. SUSHI allows for the automated retrieval of COUNTER usage reports. It is implemented as a XML-SOAP based web service.

[18] COUNTER Code of Practice Release 4: https://www.projectcounter.org/wp-content/themes/project-counter-2016/pdfs/COUNTER-code-of-practice.pdf?v=1488965556
[19] https://www.projectcounter.org/counter-release-5-code-practice/
[20] For a list of registered values for data types and metric types see: http://www.niso.org/workrooms/sushi/values/
[21] IRUS-UK Code of Practice: http://irus.mimas.ac.uk/help/toolbox/IRUS-UK_CoP_V1.0_May_2015.pdf

SUSHI-Lite[1] however is based on a RESTful interface and uses the JSON data format. It allows for the retrieval of smaller snippets of usage data for e.g. single journals and articles.[22]

A plugin implementation of SUSHI-Lite is available for the popular OJS platform. It exposes article- and journal-level metrics for reporting and analysis to aggregative services.[23]

[22] NOTE: the NISO SUSHI-Lite Working Group is currently updating the SUSHI-Lite specification to a new version based on Swagger (http://swagger.io/)

[23] The ULS at University of Pittsburgh has developed the OJS SUSHI-Lite plugin: https://github.com/ulsdevteam/ojs-sushiLite-plugin

# 2 | USAGE ANALYTICS SERVICE SPECIFICATION

OpenAIRE Usage Analytics and Usage Statistics serve as a basis for

- reliable, comparable standards-based statistics (COUNTER-conformant, bot filtering)
- reporting to stakeholders (e.g. as a repository dashboard feature)
- accumulated usage statistics of repository items which are hosted in multiple data providers (de-duplication)
- provision of usage statistics as an open metric via a standardized API (SUSHI-Lite) for 3rd party re-use.

Basically the service supports the tracking from individual data providers (repositories, e-journals, CRIS) and the collection of usage statistics reports from statistics aggregators. In addition the service also generates statistics on the usage of the OpenAIRE portal and OpenAIRE API.

The main components of the service as depicted in Figure 2 are

- the Piwik platform and OpenAIRE statistics database for the storage of usage events and usage analysis
- the Piwik tracker code and SUSHI-Lite client for gathering usage events and usage statistics respectively
- the cleaning scripts that apply rules from the COUNTER Codes of Practice on the usage data logs
- and the SUSHI-Lite API endpoint for the provision of consolidated usage statistics and OpenAIRE dashboard and portal pages for the presentation respectively.

The components interact with other parts of the OpenAIRE infrastructure, namely the repository dashboard, the data source profile management and the OpenAIRE portal.

Integration of repository usage statistics comprises the following steps:

- registration of the repository in the Piwik platform via the repository dashboard
- generation and installation of the tracking script
- tracking of usage events from the repository in Piwik
- application of COUNTER rules on usage data and storing them to the OpenAIRE statistics DB
- presentation of usage statistics on the OpenAIRE Portal and dashboard and provision via the SUSHI-Lite API endpoint.

Usage statistics reports collected from statistics aggregators are directly imported in the OpenAIRE statistics database.

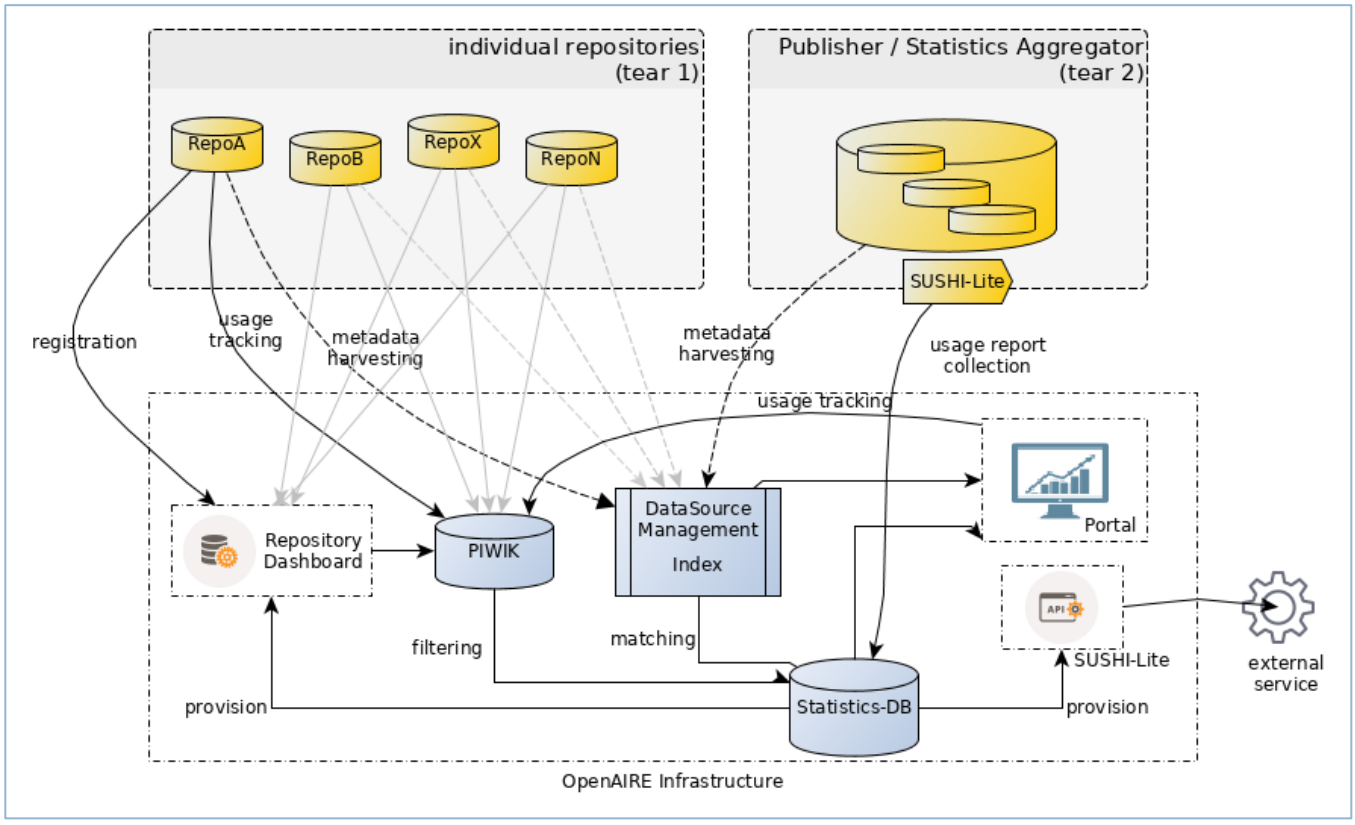


**FIGURE 2: USAGE STATISTICS SERVICE ARCHITECTURE COMPONENTS**

## 2.1 Registration of the repository in the OpenAIRE Piwik platform

The first step of the usage tracking process is the registration of the repository to the Piwik platform. This is done by the data provider manager via the repository dashboard. A unique identifier of the Piwik instance is generated and associated with the repository. The identifier is stored and accessible in the data source profile.

Once the repository is registered, the Piwik platform generates a JavaScript snippet[24] [25] which is sent to the repository manager to embed it on the repository pages for tracking. This script tracks the usage events and stores them to the Piwik platform. The tracking code looks as follows:

```
<!-- Piwik -->
<script type="text/javascript">
  var _paq = _paq || [];
  _paq.push(["trackPageView"]);
  _paq.push(["enableLinkTracking"]);
  (function() {
```

[24] https://developer.piwik.org/guides/tracking-javascript-guide
[25] what data does Piwik track: https://piwik.org/faq/general/faq_18254/

```
    var u=(("https:" == document.location.protocol) ? "https" : "http") + ":{$PIWIK_URL}/";
    _paq.push(["setTrackerUrl", u+"piwik.php"]);
    _paq.push(['setCustomVariable', 1, 'oaipmhID',"oai:<%= baseUrl >/<%=handle >", 'page']);
    _paq.push(["setSiteId", "{$IDSITE}"]);
    var d=document, g=d.createElement("script"), s=d.getElementsByTagName("script")[0];
g.type="text/javascript";
    g.defer=true; g.async=true; g.src=u+"piwik.js"; s.parentNode.insertBefore(g,s);
  })();
</script>
<noscript><p><img src="/piwik/piwik.php?idsite=13&amp;rec=1" style="border:0" alt="" /></p></noscript>
<!-- End Piwik Code -->
```

The code is executed when a page being tracked loads in a web browser. It contains all the functionality that Piwik requires in order to track a visit. {$PIWIK_URL} refers to the URL of the Piwik installation, with either a leading http:// or https://. The {$IDSITE} is the unique identifier of the repository being tracked in the Piwik platform.

## 2.2 Logging of usage events in the Piwik platform

When the URL to Piwik has been generated, the JavaScript code sends the piwik.js file from the Piwik platform in the loaded page. When the piwik.js has been delivered to the visitor's browser, the *getTracker* function creates the tracker object. A page view is triggered with the *trackPageView* function, which logs a visit to the page. Finally, the function *enableLinkTracking* is used to track clicks on links that leave the website, as well as file downloads.

The final section of the JavaScript snippet includes the Piwik's image tracker code, included in the noscript section. This code includes the piwik.php file as an image in the page which is tracked. This "fake" image allows tracking of visits from people who have JavaScript blocked.

A number of parameters are tracked by the Piwik's JavaScript. The most important ones are shown in Table 2 below:

**TABLE 2: TRACKING PARAMETERS**

| Parameter | Description |
|---|---|
| idSite | the ID of the repository |
| idVisit | a visitor/session ID (an 8 byte binary string) |
| visitIp | the IP address of the visitor |
| action | the action performed (view, download, outlink, etc) |
| url | the url of the requested item |
| timestamp | the date & time of the request |
| OAI-PMH Identifier | the Open Access Initiative identifier of the item being viewed/downloaded |

| | |
|---|---|
| agent | the Web Browser and the operating system of the visitor |
| referrer | The url that is linked to the item requested |

It should be noted that the OAI-PMH Identifier is not a default tracking value for Piwik. It has been specified using a functionality provided by Piwik, that allows the definition of custom variables ('setCustomVariable' in the JavaScript).

Usage activity records can be accessed via the Piwik API and e.g. transferred to the OpenAIRE statistics DB. An example of an API call which returns in JSON format, visitors' information for a specific day, is given below:

```
https://analytics.openaire.eu/index.php?module=API&method=Live.getLastVisitsDetails&format=JSON&idSite=14&period=day&date=yesterday&expanded=1&token_auth=123456abcdefg&filter_limit=1000
```

The parameters of the above API call and a snapshot result are given in Table 3 and Table 4:

**TABLE 3: PIWIK API RETURN PARAMETERS**

| Parameter | Description |
|---|---|
| baseURL of the piwik platform | analytics.openaire.eu/ |
| module | piwik's API |
| method called | Live.getLastVisitsDetails<br>The method returns data from the last visitors to the repository for the given period. A set of API methods that could be used is provided in the Piwik documentation. |
| idSite | the unique identifier of site (repository) being tracked |
| period | range |
| date | the requested period |
| format of the result | json |
| expanded results | 1 or 0 |
| filter limit | maximum number of retrieved records |
| token_auth | token to authenticate for API requests |

**TABLE 4: RESULT SNAPSHOT**

```
[
  {
    "idSite": "14",
    "idVisit": "6040778",
    "visitIp": "191.251.221.93",
```

```
    "visitorId": "0a7c7c8858aa0f9c",
    "actionDetails": [
      {
        "type": "action",
        "url": "https://bibliotecadigital.ipb.pt/handle/10198/8470",
        "pageTitle": "Biblioteca Digital do IPB: Estilos de liderança no terceiro setor e repercussão nos níveis
de motivação dos colaboradores",
        "pageIdAction": "1089484",
        "serverTimePretty": "Jan 21, 2017 23:57:43",
        "pageId": "35559074",
        "generationTime": "0.84s",
        "timeSpent": "1",
        "timeSpentPretty": "1s",
        "icon": null,
        "timestamp": 1485043063
      },
      {
        "type": "action",
        "url": "https://bibliotecadigital.ipb.pt/handle/10198/8470",
        "pageTitle": "Biblioteca Digital do IPB: Estilos de liderança no terceiro setor e repercussão nos níveis
de motivação dos colaboradores",
        "pageIdAction": "1089484",
        "serverTimePretty": "Jan 21, 2017 23:57:44",
        "pageId": "35559075",
        "customVariables": {
          "1": {
            "customVariablePageName1": "oaipmhID",
            "customVariablePageValue1": "oai:bibliotecadigital.ipb.pt:10198/8470"
          }
        },
        "generationTime": "0.84s",
        "timeSpent": "4",
        "timeSpentPretty": "4s",
        "icon": null,
        "timestamp": 1485043064
      },
```


## 2.3 Applying COUNTER rules on usage events

A log file in JSON format is returned by Piwik, for each day of usage activity, and processed further. An important step of this process is the cleaning of usage activity which is caused by

machines like web bots or spiders. Such software systematically browses websites in order to enhance web indexing but their activities affect usage traffic statistics since they result in the logging of non-legitimate usage activity.

To avoid such non-legitimate traffic Piwik maintains a community-contributed list of referrer spammers maintained. The list is stored in a file named *spammers.txt* and contains one (bot/spider) host per line. This list is included in each Piwik release so that referrer spam is filtered automatically. Piwik also automatically updates this list to its latest version every week.

In addition, Piwik offers a plugin, named *BotTracker* that allows the exclusion and separate tracking of Bots' activity, i.e. bots, spiders and web crawlers. The list of bots that are tracked by the specific Piwik plugin is given in the appendix. Therefore, a preliminary cleaning process has been applied to usage activity by Piwik itself and thus a large amount of bots' activity has been excluded from the log files retrieved by Piwik's REST API.

The main cleaning process in usage data-logs, is the application of the COUNTER Code of Practice rules, as described in the last version of the COUNTER framework. As stated above, the COUNTER framework provides an international, extendible Code of Practice for e-Resources that allows the usage of online information products and services to be measured in a credible, consistent and compatible way using vendor-generated data.

COUNTER specifies the following return codes and time filters for data processing [26] [27]:

a. Only successful and valid requests should be counted. For web server logs successful requests are those with specific NCSA return codes. (200 and 304). The standards for return codes are defined and maintained by NCSA. In case key events are used their definition should match the NCSA standards. (For more information see Appendix D: Guidelines for Implementation.)
b. Records generated by the server together with the requested page (e.g. images, gif's, style sheets (.css)) should be ignored.
c. All users' double-clicks on an http-link should be counted as only 1 request. The time window for occurrence of a double-click should be set at 10 seconds between the first and the second mouse-click.

There are a number of options to make sure that a double click comes from one and the same user:

1. where only the IP address of a user is logged that IP should be taken as the field to trace double-clicks
2. when a session-cookie is implemented and logged, the session-cookie should be used to trace the double-clicks.
3. when user-cookies are available and logged, the user-cookie should be used to trace double- clicks
4. when the username of a registered user is logged, this username should be used to trace double-clicks.

The options 1 to 4 above have an increasing level of reliability for filtering out double-clicks: option 1 has the lowest level of precision (and may lead to under reporting from the vendor perspective) while with option 4 the result will be optimal.

The downloading and rendering of a PDF, image, video clip or audio clip may take longer than the rendering of an HTML page. Therefore, requests by one and the same IP/username/session- or user cookie for one and the same PDF, image, video clip or audio clip should be counted as a single request if these multiple requests occur

[26] https://www.projectcounter.org/code-of-practice-sections/data-processing/#returncodesandtimefilters

[27] In COUNTER R5 draft the double-click rule is changed so that the 30 second window applies to HTML, PDF or any other format (a long HTML page can take longer to download than a short PDF!)

within a 30 seconds time window. These multiple requests may also be triggered by pressing a refresh or back button on the desktop by the user.

In OpenAIRE's case the 10s and 30s time window for views and downloads are implemented respectively. Sessions are automatically identified by Piwik.

Regarding repository records with multiple files associated (e.g. multiple book chapters), download requests are counted multiple times if the identifier (e.g. Handle, DOI) is the same.

## 2.4 Storage in the OpenAIRE statistics database

The resulting "cleaned" usage information is stored in the OpenAIRE statistics DB. Piwik logs are stored in a separate database table named *piwiklog*. A description is given in Table 5:

**TABLE 5: PIWIK LOG TABLE**

| Column Name | Data Type | Description |
|---|---|---|
| source | integer | id of the repository registered in Piwik |
| idvisit | text | session ID |
| visitip | text | IP of the visitor |
| host | text | host name of the visitor |
| country | text | visitor's country |
| action | text | view or outlink or download |
| url | text | url requested |
| entityid | text | OAIPMHID or OpenAIRE portal ID |
| entitytype | text | project or datasource or publication |
| sourceitemtype | text | repository item or portal item |
| timestamp | text | timestamp of the visit |
| referrername | text | referrer of the visit (e.g. Google scholar) |
| agent | text | agent of the visit (eg. Chrome, Safari) |

The piwiklog database table is connected to other OpenAIRE database tables using the entityid and the *orid* of each entitytype. In this manner, further information can be retrieved for usage data, such as metadata information.

The *piwiklog* table is also connected with monthly statistics table, using the id of the repository at the Piwik platform, i.e. the *source*. In this table the results of the statistical analysis are stored. In particular, for each repository and for each month of usage activity, we calculate the number of views and the number of downloads. A snapshot of the DB schema is shown in Figure 3:

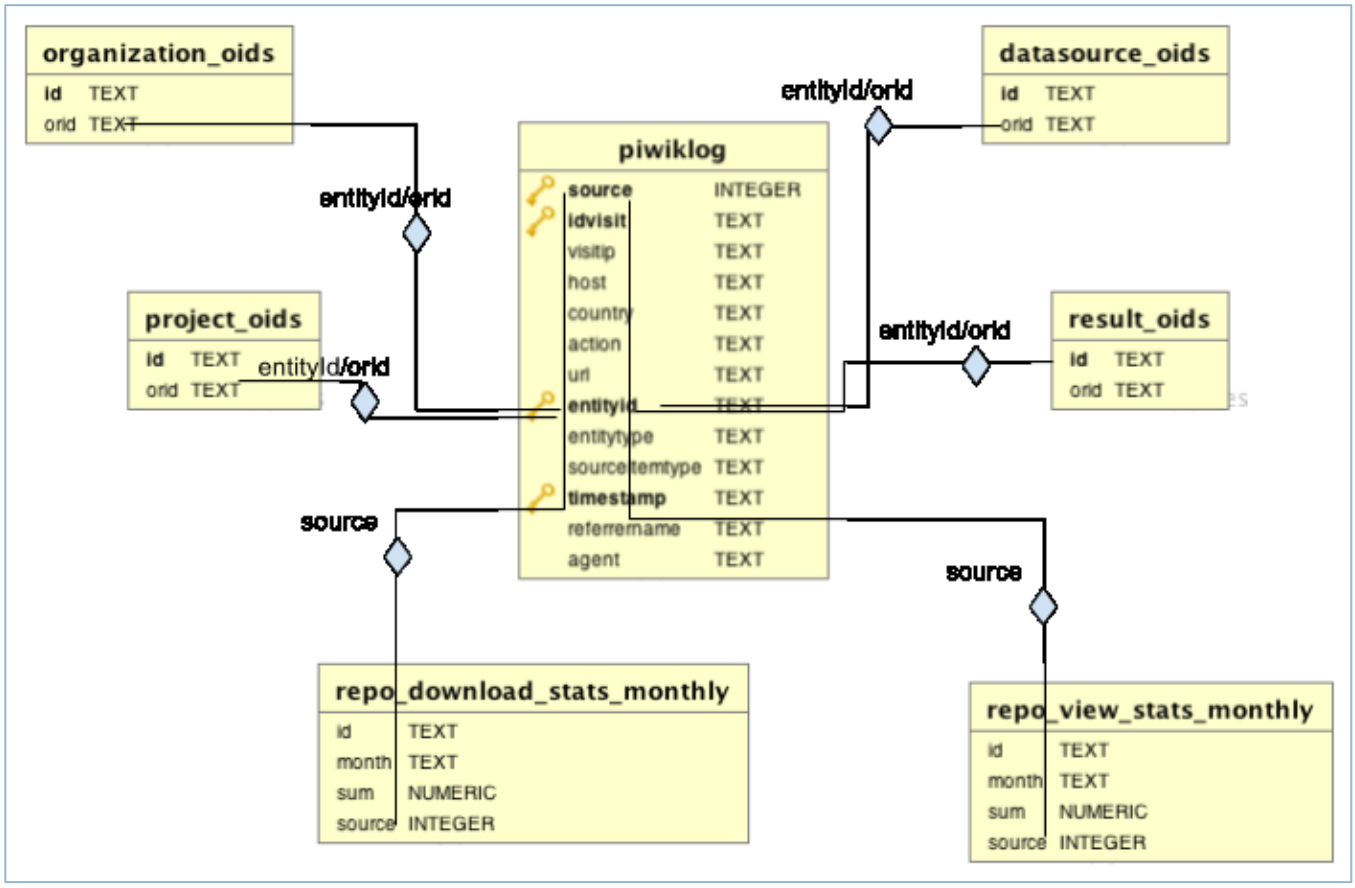


**FIGURE 3 OPENAIRE STATISTICS DB**

## 2.5 Representation and Provision of Usage Statistics

After COUNTER rules are applied, the resulting usage statistics are represented in the OpenAIRE data-provider dashboard for repository mangers and in the OpenAIRE portal per data source and document using a variety of charts, e.g. shown in Figure 4.

The statistics can be requested as COUNTER Reports. The relevant report types generated by OpenAIRE and aligned with IRUS-UK are:

- IR-1 - Item Report 1, number of successful item download requests by month and repository
- JR-1 - Journal Report 1, number of successful full-text article requests by month and journal
- RR-1 - Repository Report 1, number of successful item downloads for all repositories participating in the usage statistics service

The reports can be extended by the metrics type “abstract” to report about metadata views.

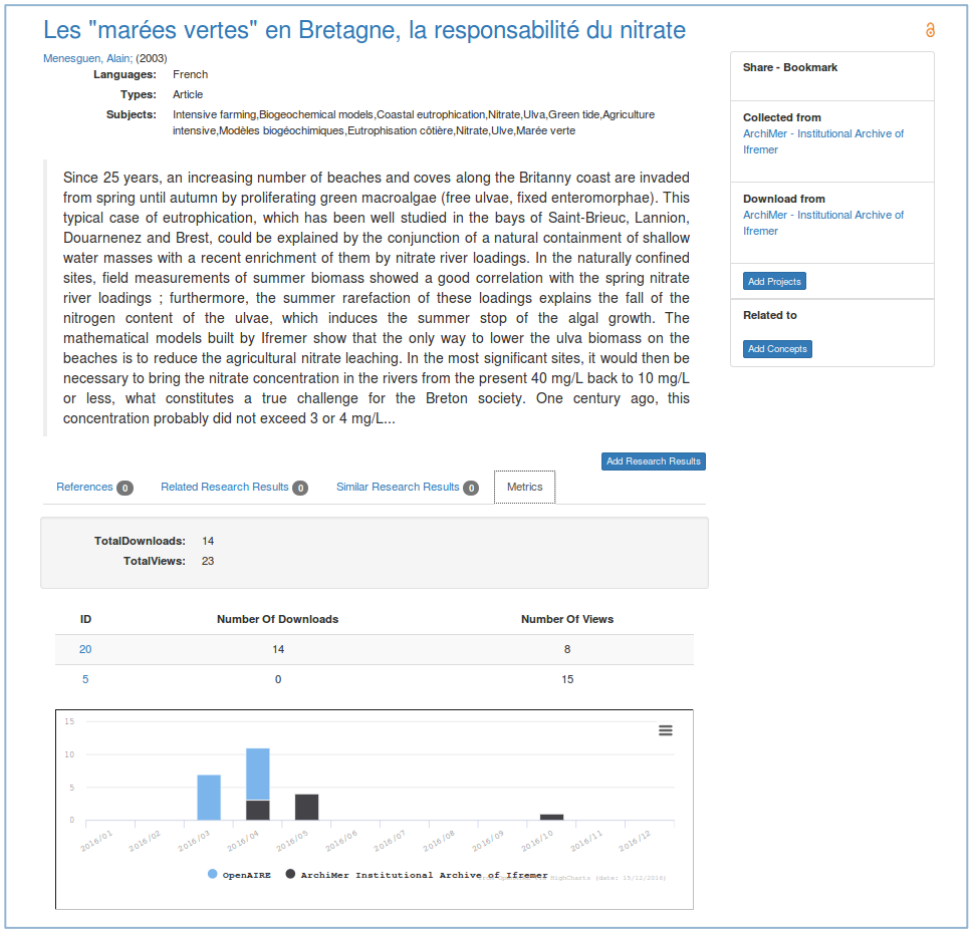


**FIGURE 4: USAGE PER PUBLICATION**

An API functionality is also provided, following the SUSHI-Lite protocol for retrieving usage statistics. An example of an API call followed by a snapshot of the result is given below:

```
http://vatopedi.di.uoa.gr:8080/stats/GetReport/?Report=IR1&Release=4&RequestorID=&BeginDate=
2016-01&EndDate=2016-10&RepositoryIdentifier=&ItemIdentifier=oid:oai:archimer.ifremer.fr:38461&
ItemDataType=&hasDOI=&Granularity=Monthly&Callback=&Pretty=Pretty
```

The result is returned in JSON.

```
{
  "ReportResponse" : {
    "@Created" : "2017-01-05 12:04:07+0200",
    "Requestor" : {
      "ID" : "anonymous"
    },
    "ReportDefinition" : {
      "@Name" : "IR1",
      "@Release" : "4",
      "Filters" : {
        "UsageDateRange" : {
          "Begin" : "2016-01",
          "End" : "2016-10"
```

```
        },
        "Filter" : [ {
          "Name" : "ItemIdentifier",
          "Value" : "oid:oai:archimer.ifremer.fr:38461"
        } ],
        "ReportAttribute" : [ {
          "Name" : "Granularity",
          "Value" : "Monthly"
        }, {
          "Name" : "ReportItemCount",
          "Value" : "1"
        } ]
      }
    },
    "Report" : {
      "Report" : {
        "@Created" : "2017-01-05 12:04:07+0200",
        "@Version" : "4",
        "@Name" : "IR1:4",
        "Vendor" : {
          "Name" : "OpenAIRE",
          "Contact" : {
            "Contact" : "OpenAIRE Helpdesk",
            "E-mail" : "helpdesk@openaire.eu"
          }
        },
        "Customer" : {
          "ID" : "anonymous",
          "ReportItems" : [ {
            "ItemIdentifier" : [ {
              "Type" : "URL",
              "Value" :
"http://archimer.ifremer.fr/doc/00273/38461/36864.pdf ;http://archimer.ifremer.fr/doc/00273/38461/ ;"
            }, {
              "Type" : "OAI",
              "Value" : "oai:archimer.ifremer.fr:38461"
            }, {
              "Type" : "OPENAIRE",
              "Value" : "od_________7::fb90de6f20d79783d05749d8f60417d5"
            } ],
            "ItemPlatform" : "",
            "ItemDataType" : "Article",
            "ItemName" : "Suivi estival des lagunes m diterran ennes fran aises   Bilan des r sultats 2014",
            "ItemPerformance" : [ {
              "Period" : {
                "Begin" : "2016-01-01",
                "End" : "2016-01-31"
              },
              "Category" : "Requests",
              "Instance" : {
                "MetricType" : "ft_total",
                "Count" : "0"
              }
            }, {
              "Period" : {
                "Begin" : "2016-02-01",
                "End" : "2016-02-29"
              },
              "Category" : "Requests",
              "Instance" : {
                "MetricType" : "ft_total",
                "Count" : "0"
              }
            }, {
              "Period" : {
                "Begin" : "2016-03-01",
                "End" : "2016-03-31"
```

```
    },
    "Category" : "Requests",
    "Instance" : {
      "MetricType" : "ft_total",
      "Count" : "0"
    }
  }, {
    "Period" : {
      "Begin" : "2016-04-01",
      "End" : "2016-04-30"
    },
    "Category" : "Requests",
    "Instance" : {
      "MetricType" : "ft_total",
      "Count" : "10"
    }
  }, {
    "Period" : {
      "Begin" : "2016-05-01",
      "End" : "2016-05-31"
    },
    "Category" : "Requests",
    "Instance" : {
      "MetricType" : "ft_total",
      "Count" : "3"
    }
  }, {
    "Period" : {
      "Begin" : "2016-06-01",
      "End" : "2016-06-30"
    },
    "Category" : "Requests",
    "Instance" : {
      "MetricType" : "ft_total",
      "Count" : "5"
    }
  }, {
    "Period" : {
      "Begin" : "2016-07-01",
      "End" : "2016-07-31"
    },
    "Category" : "Requests",
    "Instance" : {
      "MetricType" : "ft_total",
      "Count" : "10"
    }
  }, {
    "Period" : {
      "Begin" : "2016-08-01",
      "End" : "2016-08-31"
    },
    "Category" : "Requests",
    "Instance" : {
      "MetricType" : "ft_total",
      "Count" : "3"
    }
  }, {
    "Period" : {
      "Begin" : "2016-09-01",
      "End" : "2016-09-30"
    },
    "Category" : "Requests",
    "Instance" : {
      "MetricType" : "ft_total",
      "Count" : "3"
    }
  }, {
```


```
            "Period" : {
              "Begin" : "2016-10-01",
              "End" : "2016-10-31"
            },
            "Category" : "Requests",
            "Instance" : {
              "MetricType" : "ft_total",
              "Count" : "3"
            }
          } ]
        } ]
      }
    }
  }
}
```


## 2.6 Piwik reporting scheme

Piwik offers a real time reporting scheme, where the platform administrators can view usage statistics in the tracked repository. The same information can be also provided to repository administrators, using specific credentials. The following snapshots in Figure 5 and Figure 6 show types of usage activity as displayed in the Piwik platform.


**FIGURE 5: REAL TIME USAGE ACTIVITY**

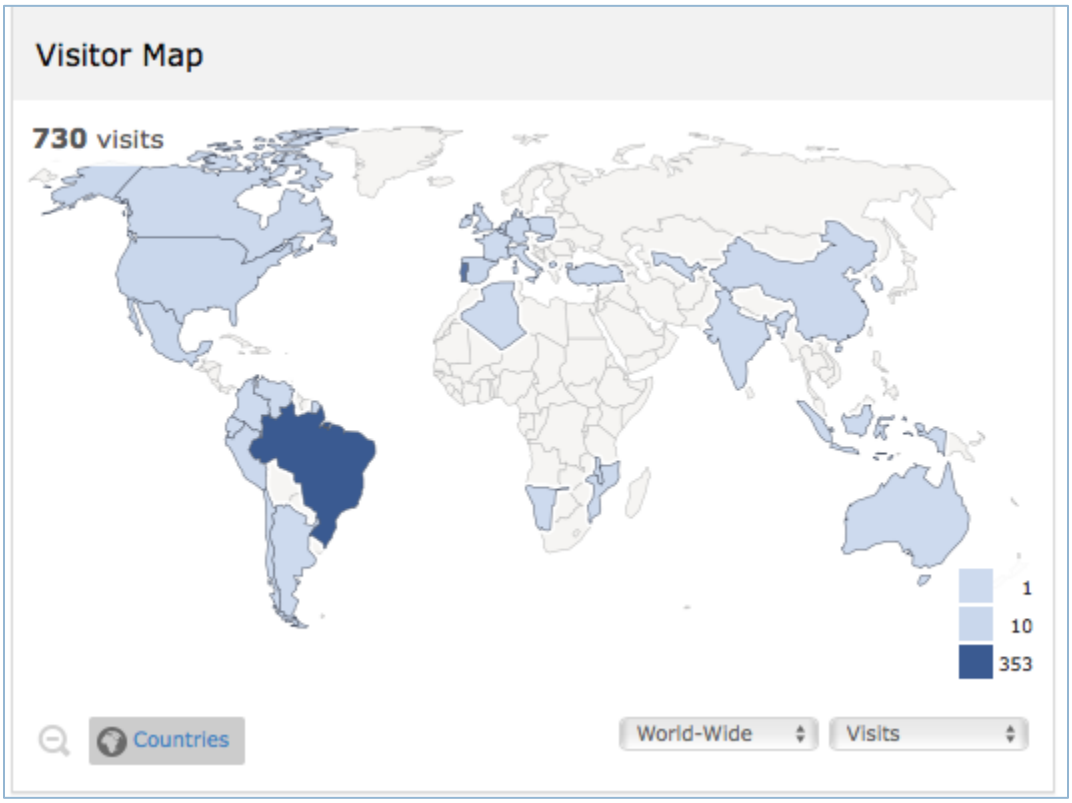


**FIGURE 6: REAL TIME VISITOR MAP**

## 2.7 Scalability of Piwik

The Piwik software is designed to scale up to millions of pageviews per month. Its only limitation might be merely the underlying hardware.

Piwik in OpenAIRE tracked a total of 1.3 million usage events (already cleaned from bots) from 35 websites (OpenAIRE portal and repositories registered in Piwik) in December 2016.

OpenAIRE counts roughly 800 data providers (12/2016). While the number of registered websites (e.g. data providers) in Piwik is not limited, the hardware environment where Piwik is installed needs to be carefully configured so that the number of tracked usage events scales well[28].

If the tracking API becomes a bottleneck, Piwik can be optimized by queued tracking in Redis so that usage events get cached before they are written to the Piwik database[29] [30] [31].

## 2.8 Gathering Usage Reports from Data Providers supporting the SUSHI-Lite protocol

[28] https://piwik.org/docs/data-limits/
[29] How to configure Piwik Tracking for high reliability: https://piwik.org/faq/how-to/faq_17033/
[30] Queued tracking: https://plugins.piwik.org/QueuedTracking
[31] Optimize and scale Piwik: https://piwik.org/docs/optimize/

OpenAIRE's Usage Statistics Service can also collect COUNTER usage reports from data providers supporting the SUSHI-Lite protocol. This is currently supported by the IRUS-UK aggregation service and available for OJS platforms. The SUSHI-Lite API is exploited to retrieve aggregated statistics as shown in the following example:

```
http://irus.mimas.ac.uk/api/sushilite/v1_7/GetReport/?Report=AR4&Release=4&RequestorID=Jisc&BeginDate=2016-01&EndDate=2016-03&RepositoryIdentifier=irusuk%3A28&ItemIdentifier=&hasDOI=&Granularity=Monthly&Callback=&Pretty=Pretty
```

A snapshot of the returned results is given below:


```
{
	"ReportResponse": {
		"@Created": "2017-01-23 13:47:44+00:00",
		"Requestor": {
			"ID": "Jisc"
		},
		"ReportDefinition": {
			"@Name": "AR4",
			"@Release": "4",
			"Filters": {
				"UsageDateRange": {
					"Begin": "2016-01",
					"End": "2016-03"
				},
				"Filter": [
					{
						"Name": "RepositoryIdentifier",
						"Value": "irusuk:28"
					}
				],
				"ReportAttribute": [
					{
						"Name": "Granularity",
						"Value": "Monthly"
					},
					{
						"Name": "ReportItemCount",
						"Value": 3164
					}
				]
			}
		},
		"Report": {
			"Report": {
				"@Created": "2017-01-23 13:47:44+00:00",
				"@Version": "4",
				"@Name": "AR4:4",
				"Vendor": {
					"Name": "IRUS-UK",
					"Contact": {
						"Contact": "IRUS-UK Helpdesk",
						"E-mail": "irus.mimas.ac.uk"
					}
				},
				"Customer": {
					"ID": "Jisc",
					"ReportItems": [
						{
							"ItemIdentifier": [
								{
									"Type": "URL",
```

```
                                                        "Value":
"http:\/\/eprints.ecs.soton.ac.uk\/13310\/"
                                                },
                                                {
                                                        "Type": "DOI",
                                                        "Value": "10.1007\/s10550-006-0029-6"
                                                },
```


IRUS-UK aggregated statistics are not processed and stored directly in the statistics table.

# 3 | RESULTS FROM PILOT IMPLEMENTATION

The first step towards the implementation of the usage analytics service was running a pilot with a number of repositories who participate in OpenAIRE. The pilot started with the following three portugese repositories:

- 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 7 presents the metadata views and downloads of the articles of the tracked repositories, whilst Figure 8, we present the accumulated information of same publications but tracked from two different repositories.

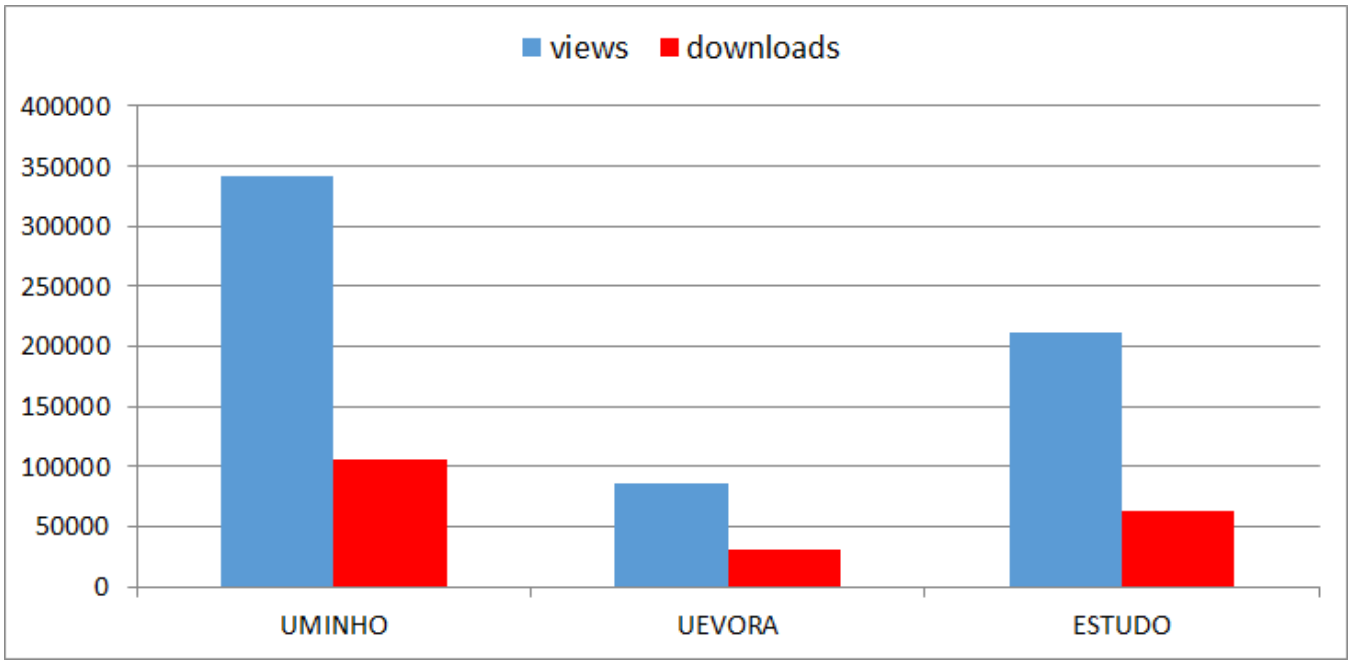


FIGURE 7: NUMBER OF VIEWS AND DOWNLOADS OF TRACKED PUBLICATION ON REPOSITORY LEVEL

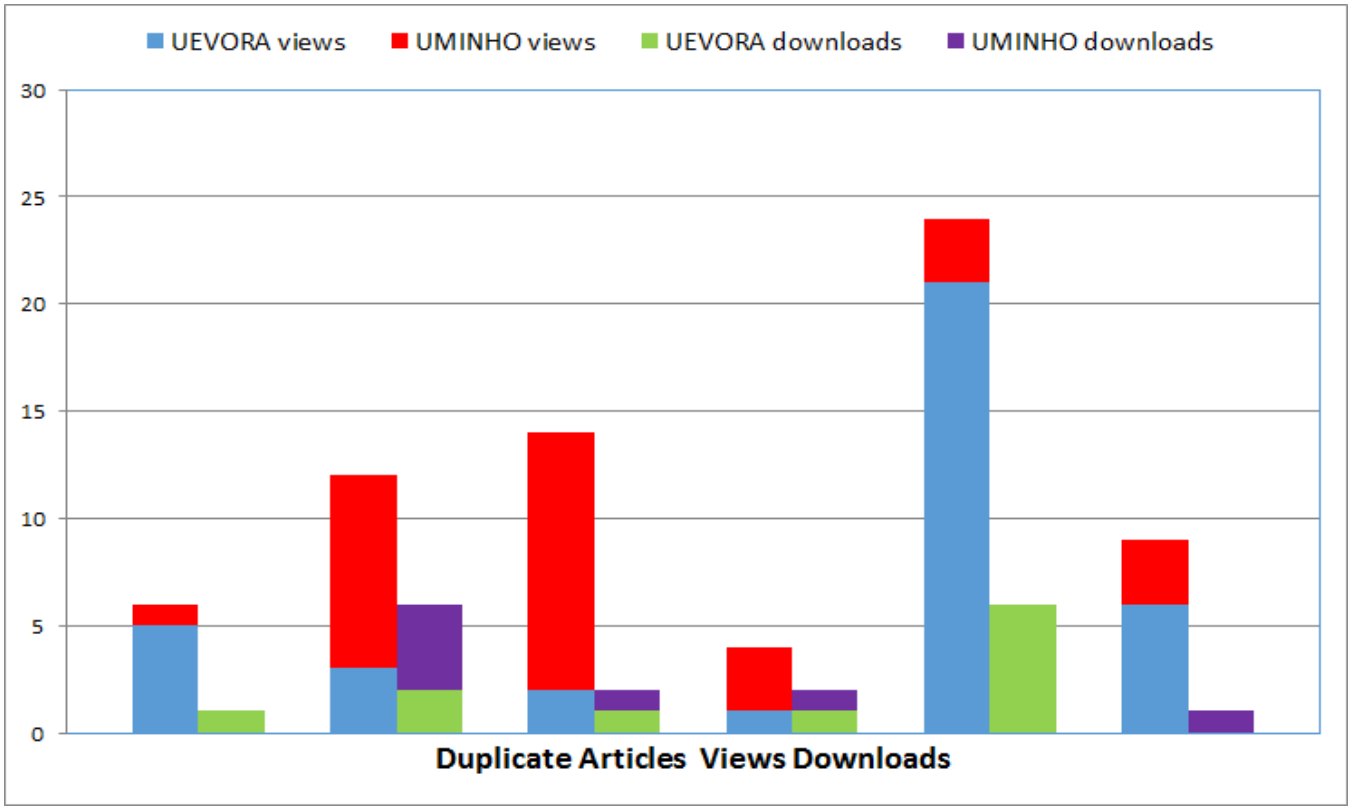


**FIGURE 8: ACCUMULATED NUMBER OF VIEWS AND DOWNLOADS OF TRACKED PUBLICATIONS HOSTED IN DIFFERENT REPOSITORIES**

Currently, more repositories are registered in OpenAIRE's Piwik platform. The table in the appendix lists the repositories registered in Piwik and the day that they have been initially registered, whilst in Figure 9 we present the usage analytics, i.e., repository views and downloads for a period of a year from 1st January 2016 to 10th December 2016.

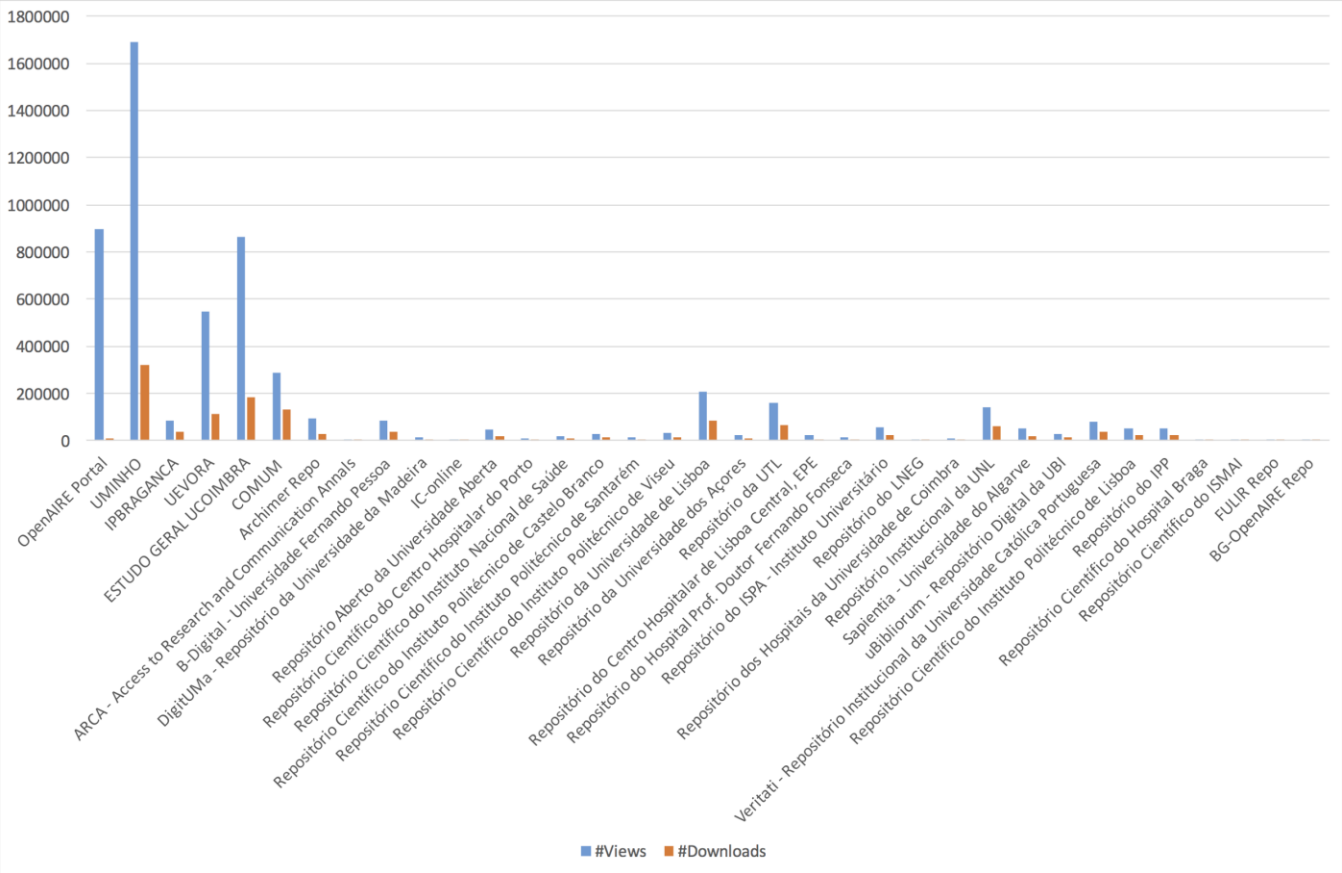


**FIGURE 9: VIEWS AND DOWNLOADS IN THE OPENAIRE PORTAL AND TRACKED REPOSITORIES PARTICIPATING IN THE PILOT FOR THE PERIOD 01-JAN-2016 – 10-DEC-2016**

## 3.1 Comparison of usage activity between Piwik and Google Analytics

Two different types of usage activity were analyzed:

1. Usage activity on the OpenAIRE portal.
2. Usage activity on a repository.

1. OpenAIRE Portal Usage Activity

Logs from Piwik and Apache server were examined for two different days, 02 September 2016 and 13 September 2016. Piwik logs were retrieved using the Piwik API.

A simple preprocessing was applied to Apache logs as follows:

- records containing the words relevant to OpenAIRE items, such as “articleId”, “datasource” and etc, were kept.
- records containing the terms “bot”, “slurp” (yahoo bot) and “spider” were removed
- records with error HTTP error codes (e.g 404) were removed

Piwik log was already cleared from known bots, using the BotTracker plugin. The results of the preprocessing phase are shown in Table 6 below:

**TABLE 6: LOG RESULTS**

| Log Date | No of Apache records | No of Apache records after preprocessing | No of Apache records NOT found in Piwik |
|---|---|---|---|
| 02-09-2016 | 171.000 | 8.666 | 4.272 |
| 13-09-2016 | 179.250 | 9.135 | 4.104 |

From the above we notice that 50%-60% of usage activity recorded in Apache was not recorded in Piwik. Further analysis of the Apache records that had not been tracked by Piwik is shown in Table 7:

**TABLE 7: APACHE LOG RECORDS NOT TRACKED BY PIWIK**

| Log Date | Number of entries | Status |
|---|---|---|
| 02-09-2016 | 3.292 | Referrer: "-" |
| | 296 | Referrer is from worldwidescience.org |
| 13-09-2016 | 2.391 | Referrer: "-" |
| | 416 | Referrer is from worldwidescience.org |

The majority of non-tracked usage activity in Piwik was recorded with an empty, aka "-" at the referrer field. This might belong to visitors behind firewalls or other software that does not send referrer information. However, there were records with no referrer information which were correctly tracked by Piwik. There a number of records from "worldwidescience.org" that have not been recorded.

2. Repository (IPBRAGANCA) Usage Activity

Usage activity from a Portuguese repository participating in the usage statistic pilots was examined. In particular, Apache logs and piwik logs from the IPBraganca repository were compared, for the 4th October 2016. Similar to the OpenAIRE portal log, simple, i.e. no COUNTER rules, preprocessing approach was applied and only the pdf records were kept from the logs.

The results of the preprocessing phase are shown in Table 8:

**TABLE 8: LOG RESULTS AFTER PREPROCESSING**

| Log type | Number of full-text requests |
|---|---|

| | |
|---|---|
| Apache | 4.197 |
| Piwik | 388 |

Less than 10% of the Apache log entries were also recorded in Piwik. The analysis shown that the majority of Apache log entries are not tracked by Piwik, 2.003 log entries in particular, had some google search engine in the referrer field, such as google.com, google.pt, etc. Some of these entries (around 900) were usually followed by an entry were the referrer field is the same page with the one requested. This activity was also not tracked by Piwik. An example of such log entries taken from the Apache log file, are given below:

187.15.5.145 - - [04/Oct/2016:00:10:54 +0100] GET /bitstream/10198/3961/1/O%20estudo%20de%20caso%20como%20estrat%C3%A9gia%20de%20investiga%C3%A7%C3%A3o%20em%20educa%C3%A7%C3%A3o.pdf HTTP/1.1 200 24552 https://bibliotecadigital.ipb.pt/bitstream/10198/3961/1/O%20estudo%20de%20caso%20como%20estrat%C3%A9gia%20de%20investiga%C3%A7%C3%A3o%20em%20educa%C3%A7%C3%A3o.pdf Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/53.0.2785.143 Safari/537.36 383369 https://bibliotecadigital.ipb.pt/bitstream/10198/3961/1/O%20estudo%20de%20caso%20como%20estrat%C3%A9gia%20de%20investiga%C3%A7%C3%A3o%20em%20educa%C3%A7%C3%A3o.pdf

187.15.5.145 - - [04/Oct/2016:00:10:54 +0100] GET /bitstream/10198/3961/1/O%20estudo%20de%20caso%20como%20estrat%C3%A9gia%20de%20investiga%C3%A7%C3%A3o%20em%20educa%C3%A7%C3%A3o.pdf HTTP/1.1 200 266269 https://www.google.com.br/ Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/53.0.2785.143 Safari/537.36 2346101 https://www.google.com.br/

Similar to the OpenAIRE portal logs, a number of records, 1294 records in particular, with no agent information, i.e. "-" in the Agent field, were not recorded by Piwik.

In this report, the focus was on the comparison between Piwik and Apache logs. Logs, from Google Analytics had not been examined at this phase. Moreover, as we can see from the snapshots below, Google Analytics and Piwik had almost similar tracking behavior.


**FIGURE 10: USAGE ACTIVITY IN PIWIK**

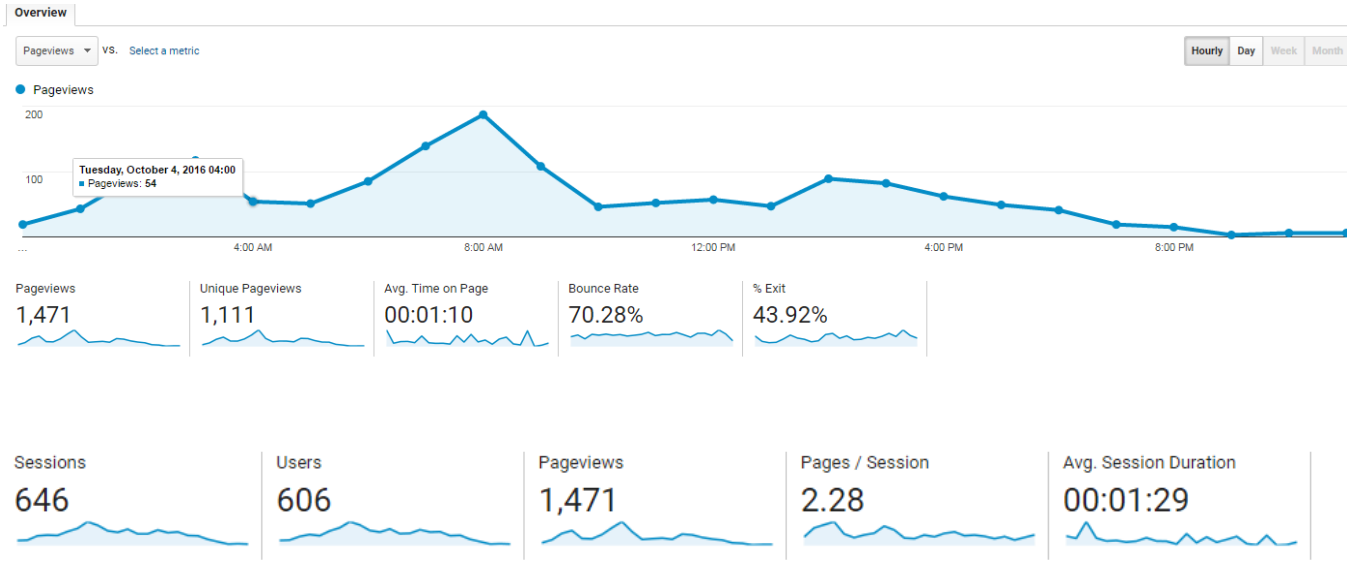


**FIGURE 11: USAGE ACTIVITY IN GOOGLE ANALYTICS**

# 4 | DISCUSSION AND NEXT STEPS

In the description about the tracking mechanism we mentioned that the Piwik JavaScript code snippet is a generic approach that can be installed in any type of repository without further development. It suffers though, from a number of shortcomings, such as blocking of tracking by Ad blockers, or missing of download requests that bypass the client software, such as the web browser.

In order to alleviate these drawbacks, we have decided to develop a new approach that will operate with the tracking code on the server's side in contrast to embedding a tracking snippet in the landing page. This approach will allow tracking of usage activity directly at the repository's backend, by monitoring the HTTP/HTTPS requests. More particularly, each time a request is sent to the repository, the Piwik tracker, e.g. developed as a Java servlet, would be triggered and an event would be sent to the Piwik platform. This is a similar approach to the IRUS-UK service, but it would exploit all the Piwik platform advantages.

The difficulty of this approach is that we have to develop different versions, maybe in different programming languages, since there is a variety of repository platforms, eg DSpace, Fedora, EPrints, even proprietary approaches like the Archimer repository. Fortunately, Piwik offers its tracker software for a number of languages, like PHP, Java etc. A minor issue is the tracking of the OAI-PMH identifier. In the JavaScript approach, this identifier is specified automatically in Piwik's custom variable, inside the script. If a server side solution is followed, this custom variable should be created on running time, using the URL of the request and the OAI-PMH preamble of the repository.

# 5 | REFERENCES

[1] SUSHI-Lite Technical Report Working Group. (n.d.). SUSHI-Lite. Retrieved December 16, 2016, from http://www.niso.org/workrooms/sushi/sushi_lite/

[2] DFG Project "Open Access Statistics" and DINI Working Group "Electronic Publishing. (2013). "Standardised Usage Statistics for Open Access Repositories and Publication Services", http://edoc.hu-berlin.de/series/dini-schriften/2013-13-en/PDF/13-oas-en.pdf

[3] IRUS-UK Service Level Agreement. (2016). http://irus.mimas.ac.uk/help/toolbox/IRUS-UK_SLA_V1.1_July_2016.pdf

[4] IRUS-UK Code of Practice v1.0. (2015). http://irus.mimas.ac.uk/help/toolbox/IRUS-UK_CoP_V1.0_May_2015.pdf

[5] COUNTER Code of Practice for e-Resources Release 4.(first published 2012). https://www.projectcounter.org/wp-content/themes/project-counter-2016/pdfs/COUNTER-code-of-practice.pdf?v=1488965556

[6] Piwik PRO.(2016). PIWIK vs. Google Analytics whitepaper, https://piwik.pro/2016/05/infographic-piwik-vs-google-analytics/

# 6 | APPENDIX

## 6.1 Participating repositories in OpenAIRE's Piwik

**TABLE 9: PARTICIPATING PILOT REPOSITORIES IN USAGE STATISTICS**

| Repository Name | Registered Date |
|---|---|
| OpenAIRE Portal | 28-11-2013 |
| UMINHO Pilot | 19-5-2015 |
| IPBRAGANCA Pilot | 6-8-2015 |
| UEVORA Pilot | 6-8-2015 |
| ESTUDO GERAL UCOIMBRA Pilot | 6-8-2015 |
| COMUM Pilot | 19-1-2016 |
| Archimer Repo | 29-3-2016 |
| ARCA - Access to Research and Communication Annals Pilot | 10-5-2016 |
| B-Digital - Universidade Fernando Pessoa Pilot | 10-5-2016 |
| DigitUMa - Repositório da Universidade da Madeira Pilot | 10-5-2016 |
| IC-online Pilot | 10-5-2016 |
| Repositório Aberto da Universidade Aberta Pilot | 10-5-2016 |
| Repositório Científico do Centro Hospitalar do Porto Pilot | 10-5-2016 |
| Repositório Científico do Instituto Nacional de Saúde Pilot | 10-5-2016 |
| Repositório Científico do Instituto Politécnico de Castelo Branco Pilot | 10-5-2016 |
| Repositório Científico do Instituto Politécnico de Santarém Pilot | 10-5-2016 |
| Repositório Científico do Instituto Politécnico de Viseu Pilot | 10-5-2016 |
| Repositório da Universidade de Lisboa Pilot | 10-5-2016 |
| Repositório da Universidade dos Açores Pilot | 10-5-2016 |
| Repositório da UTL Pilot | 10-5-2016 |
| Repositório do Centro Hospitalar de Lisboa Central, EPE Pilot | 10-5-2016 |
| Repositório do Hospital Prof. Doutor Fernando Fonseca Pilot | 10-5-2016 |
| Repositório do ISPA - Instituto Universitário Pilot | 10-5-2016 |
| Repositório do LNEG Pilot | 10-5-2016 |
| Repositório dos Hospitais da Universidade de Coimbra Pilot | 10-5-2016 |
| Repositório Institucional da UNL Pilot | 10-5-2016 |
| Sapientia - Universidade do Algarve Pilot | 10-5-2016 |
| uBibliorum - Repositório Digital da UBI Pilot | 10-5-2016 |
| Veritati - Repositório Institucional da Universidade Católica Portuguesa Pilot | 10-5-2016 |
| Repositório Científico do Instituto Politécnico de Lisboa Pilot | 10-5-2016 |
| Repositório do IPP (Recipp) Pilot | 10-5-2016 |

| | |
|---|---|
| Repositório Científico do Hospital Braga Pilot | 10-5-2016 |
| Repositório Científico do ISMAI Pilot | 10-5-2016 |
| FULIR Repo | 13-5-2016 |
| BG-OpenAIRE Repo | 13-5-2016 |

## 6.2 List of bots

Piwik offers a plugin, named *BotTracker* (https://plugins.piwik.org/BotTracker) that allows the exclusion and separately tracking bots' and spiders' activity. The list of Bots that are tracked by the specific Piwik plugin is given in the following Table:

| | |
|---|---|
| FR-Crawler | SiteBot |
| Digincore | Yandex ? |
| XoviBot | Yandex |
| Ssearch Crawler | Wget |
| Domnutch-Bot | Lycos |
| RogerBot | Infoseek 2 |
| CMS Crawler | Infoseek 1 |
| Nutch Crawler | Altavista 4 |
| Magpie Crawler | Altavista 3 |
| TurnitinBot | Altavista 2 |
| NetSeer Crawler | Altavista 1 |
| MJ12Bot | Yahoo! Slurp |
| AhrefsBot | YahooSeeker |
| MSN Bot Media | Ezooms |
| Exabot | Baiduspider |
| GoogleBot | Media Partners GoogleBot |
| Bingbot | Google Instant |
| MSN Search | |

## 6.3 OpenAIRE Guidelines for Collecting Usage Events and Provision of Usage Statistics v1

### 6.3.1 Purpose

The guidelines are aimed to provide orientation for data source managers about participation in the OpenAIRE Usage Statistics Service and about the methods and standards used to collect and process usage data in order to generate comparable, standards-based usage statistics. The guidelines follow the Release 4 of the COUNTER Code of Practice for e-Resources[32] supplemented by the IRUS-UK Code of Practice[33].

### 6.3.2 Scope of Application

The OpenAIRE Usage Statistics Service gathers usage data and consolidated usage statistics reports respectively from its distributed network of data providers (repositories, e-journals, CRIS) by utilizing open standards and protocols and exploiting reliable, consolidated and comparable usage metrics like counts of item downloads and metadata views conformant to COUNTER.

### 6.3.3 Usage Data Collection, Processing and Reporting

The OpenAIRE Usage Statistics Service supports two different ways of obtaining usage information from data providers:

1) Usage events on items in data sources, such as document landing pages and full-text files, can be tracked and will be "pushed" as raw usage data to the OpenAIRE Analytics service which is based on the open source software Piwik. The cleaning process consists of two steps. At first usage data resulting from bots, spiders and web crawlers are excluded by applying a community-maintained robot list. Subsequently the data processing rules from the COUNTER Code of Practice are applied to identify successful and valid requests, sessions, and to eliminate double clicks.
   Tracked usage events contain the following paramters:

| Parameter | Description |
|---|---|
| idSite | the ID of the repository |
| idVisit | a visitor/session ID (an 8 byte binary string) |
| visitIp | the IP address of the visitor |
| action | the action performed (view, download, outlink, etc) |
| url | the url of the requested item |
| timestamp | the date & time of the request |
| OAI-PMH Identifier | the Open Access Initiative identifier of the item being viewed/downloaded |
| agent | the Web Browser and the operating system of the visitor |

[32] https://www.projectcounter.org/code-of-practice-sections/general-information/
[33] http://irus.mimas.ac.uk/help/toolbox/IRUS-UK_CoP_V1.0_May_2015.pdf

| | |
|---|---|
| referrer | The url that is linked to the item requested |

The IP address will be anonymized in Piwik by masking the last two bytes.

2) Alternatively COUNTER conformant usage statistics can be collected from data provider endpoints that support the RESTful Standardized Usage Statistics Harvesting Initiative (SUSHI) protocol[34] [35].

Consolidated Usage Statistics will be made available in OpenAIRE by the following methods:

- In the data provider dashboard for data source managers with support of configuration updates, visualization and reporting functionalities on the data provider level
- In the OpenAIRE portal by public usage statistics visualization and reporting functionalities on the data provider and individual item level
- In the OpenAIRE-API with access to usage statistics COUNTER conformant reports supporting the RESTful SUSHI protocol.

The usage statistics made available in the OpenAIRE portal and exposed via the OpenAIRE-API are released under CC0 license.

The OpenAIRE Usage Statistics Service generates usage reports conformant to COUNTER in csv/tsv and JSON format:

- IR-1 - Item Report 1, number of successful item download requests by month and repository
- JR-1 - Journal Report 1, number of successful full-text article requests by month and journal
- RR-1 - Repository Report 1, number of successful item downloads for all repositories participating in the usage statistics service

Where possible the reports will also provide the metadata views.

### 6.3.4 Participation and Workflow

The usage statistics workflow for data providers is as follows:

- Data provider managers who wishes to participate in OpenAIRE Usage Statistics can do so in the metrics section of the data provider dashboard. The options provided for the usage data transfer mechanism are:
    1. by usage data tracking (recommended for most repositories, CRIS)
    2. by usage data reporting (recommended for national repository statistics services; publishers of e-journals)
- In case of 1) the data provider manager is provided with a unique identifier which is the websiteId in Piwik. Information about generic and platform dependent tracking plugins are provided.
- In case of 2) the data provider manager informs OpenAIRE about the SUSHI-Lite endpoint URL from where usage data reports can be queried and downloaded.
- OpenAIRE usage statistics service tracks or downloads usage data, performs cleaning operations, applies COUNTER rules, associates with corresponding metadata records in the OpenAIRE index, and generates statistics.
- The data provider manager can access usage statistics in the dashboard.

[34] About the SUSHI-standard: http://www.niso.org/workrooms/sushi
[35] About RESTful SUSHI-Lite: http://www.niso.org/workrooms/sushi/sushi_lite/

- Usage statistics is presented along with the publication metadata in the OpenAIRE portal.
- Usage statistics is exposed via the OpenAIRE SUSHI-Lite API endpoint to 3rd party services.

### 6.3.5 Responsibilities

Before data providers can officially participate in OpenAIRE Usage Statistics the implementation and configuration of the tracker plugins and SUSHI-Lite endpoints respectively will be tested and validated between the data provider manager and the OpenAIRE support.

Data Provider Managers

They keep the configuration information regarding usage statistics in the data provider dashboard up to date.

In case of platform software updates (e.g. by migration, new releases) which may affect existing metadata record identifiers the data provider manager informs the OpenAIRE support.

OpenAIRE

The OpenAIRE Usage Statistics service complies with the EU data protection directive[36] which will be replaced by the General Data Protection Regulation in 2018[37] with regard to the protection of personally identifiable information.

The Usage Statistics Service will generate visualizations and reports on a monthly basis. In order to conform to the COUNTER Code of Practice the data processing and cleaning rules will be maintained to comply with the latest release of COUNTER.

### 6.3.6 Software Support

Information about tracker and SUSHI plugins are provided at: https://github.com/openaire/usage-statistics .

Initially tracker plugins for DSpace and EPrints repositories are available as well as an SUSHI-Lite plugin for OJS.

The support of further platforms will be extended.

### 6.3.7 Maintenance of the Guidelines

The guidelines are continuously reviewed for their validity and will be updated with regards to new releases of standards for the recording and exchange of usage statistics and new releases of OpenAIRE services that record, process, represent or expose usage statistics. Participating data provider managers will be informed accordingly.

[36] officially Directive 95/46/EC: http://eur-lex.europa.eu/legal-content/EN/LSU/?uri=celex:31995L0046
[37] http://ec.europa.eu/justice/data-protection/