Ref. Ares(2021)2257140 - 01/04/2021

D3.2 - Roadmap to deliver uniform packaging of OpenAIRE-Nexus portfolio

MAR. 2021

Scholarly Communication Services for EOSC users

D3.2 – Roadmap to deliver uniform packaging of OpenAIRE-Nexus portfolio

Version 2.0 –Final
PUBLIC

This deliverable describes the roadmap to deliver the expected bi-lateral integrations/interactions of OpenAIRE-Nexus services towards the delivery of a uniform and consistent portfolio and catalogue to the EOSC.

H2020-INFRAEOSC-2020-2
Grant Agreement 101017452

Document Description

D3.2 – Roadmap to deliver uniform packaging of OpenAIRE-Nexus portfolio	
WP3 - Integration with the EOSC to provide Virtual Access	
WP participating organizations: **CNR**, UNIBI, CERN, AthenRC, UGOE, ICM, UNIBO, CNRS, CITE	
Contractual Delivery Date: 31/3/2021	Actual Delivery Date: 31/03/2021
Nature: Report	Version: 2.0 (Final)
Public	

Preparation Slip

	Name	Organisation	Date
From	Paolo Manghi	OpenAIRE AMKE	31/3/2021
Edited by	Paolo Manghi Raphael Tournoy Dimitris Pierrakos Silvio Peroni Andreas Czerniak Jose Benito Gonzalez	OpenAIRE AMKE CNRS AthenaRC UNIBO UNIBI CERN	26/3/2021
Reviewed by	Miriam Baglioni	CNR	31/3/2021
Approved by	Paolo Manghi	OpenAIRE AMKE	31/3/2021
For delivery	Eleni Koulocheri	OpenAIRE AMKE	1/4/2021

Revision History

Issue	Item	Reason for Change	Author	Organization
V1.0	Draft version 23/01/2021	Editing	Paolo Manghi	OpenAIRE AMKE

V1.1	Draft version 26/01/2021	Editing	Raphael Tournoy Dimitris Pierrakos Silvio Peroni Andreas Czerniak Jose Benito Gonzalez	CNRS AthenaRC UNIBO UNIBI CERN
V1.1	Reviewed version 31/03/2021	Feedback from review	Miriam Baglioni	CNR
V2.0	Finalizing 31/03/2021	Final drafting	Paolo Manghi	OpenAIRE AMKE

Contents

Table of Figures

Disclaimer

This document contains description of the OpenAIRE-Nexus 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 OpenAIRE-Nexus consortium and can in no way be taken to reflect the views of the European Union.

OpenAIRE-Nexus is a project funded by the European Union (Grant Agreement No 101017452).

Publishable Summary

PLANNING THE INTEGRATION WITH THE EOSC – M1-M13

This deliverable describes the actions that OpenAIRE-Nexus services will be carrying out as part of Task 3.3, in order to integrate the newly hosted services episciences.org, OpenAPC, and OpenCitations with the OpenAIRE services. The aim is to offer a coherent and uniform service catalogue (and underlying portfolio) to the EOSC.

1 INTRODUCTION

The OpenAIRE Nexus portfolio consist of the three sub-portfolios PUBLISH, MONITOR, and DISCOVER, as described in Figure 1. The portfolio welcomes into the OpenAIRE portfolio three known scholarly communications services, namely:

- **Episciences.org**: Enables management of OA journals exploiting the network of open access repositories (e.g. HAL, arXiv, Zenodo).
- **OpenAPC**: Keeping track and providing access to the Open Access record of European expenditure for article processing charges (APC) across the countries
- **OpenCitations**: Keeping track and providing Open APIs to access to the largest collection of OA (CC0) citations between DOIs as exposed by publishers world-wide.

Such services were picked among many candidates based on to several factors, including sustainability, community trust, and finally gap-analysis with respect to the OpenAIRE's offer. Each of them brings an added value to the OpenAIRE mission and therefore for EOSC users. To this aim, adding their entries to the OpenAIRE catalogue (M4.1, M5.1, M6.1), hence to the EOSC catalogue, is only a minor, although essential step. The added value comes from the technical integration of such services with the OpenAIRE portfolio services, in order for both parties to enhance their offer and effectively become the components of a uniform service portfolio.

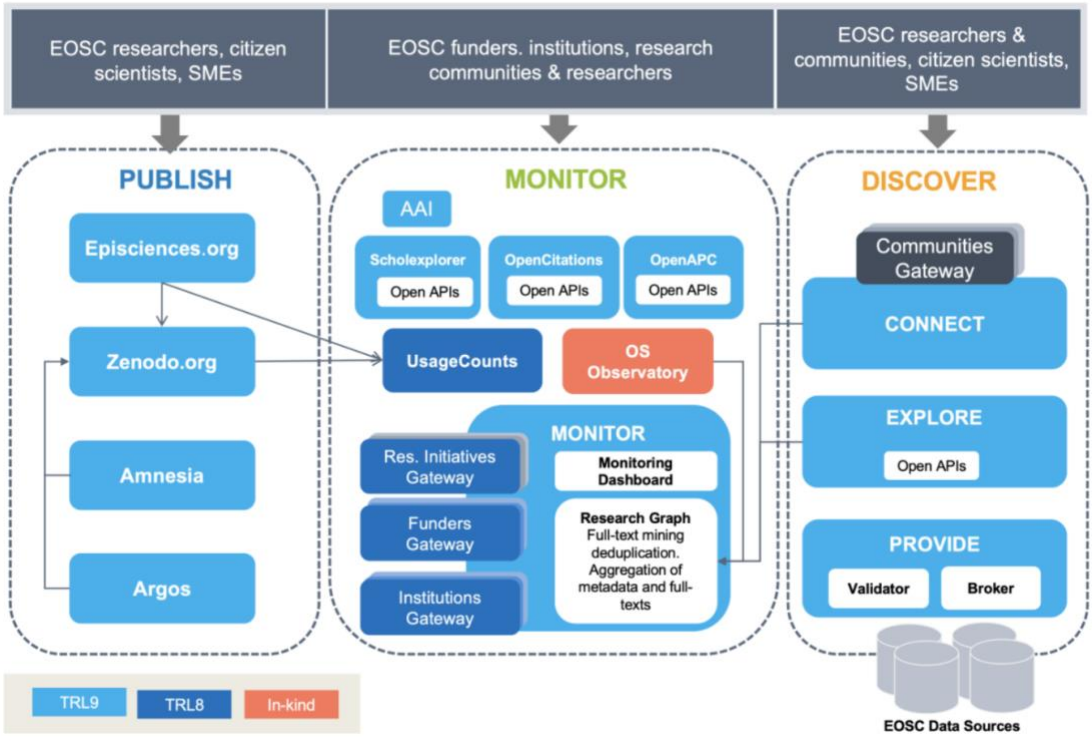


Figure 1 - OpenAIRE-Nexus service portfolio

2 EPISCIENCES.ORG

Episcience.org is involved in four activities:

- Activity 3.3.1 OpenAIRE UsageCounts to integrate Zenodo.org and episciences.org;
- Activity 3.3.2 episciences.org to integrate Zenodo.org APIs for deposition of pre-prints and vice versa, to enable submission of pre-prints to journals via Zenodo.org
- Activity 3.3.3 episciences.org to integrate with OpenAIRE PROVIDE to receive Broker notifications
- Activity 3.3.4 episciences.org to integrate with Scholexplorer/OpenCitations API

Activity 3.3.1 OpenAIRE UsageCounts to integrate Zenodo.org and episciences.org

OpenAIRE UsageCounts Service is the Usage Statistics service of OpenAIRE Research Graph and is part of the Monitor Portfolio of services. UsageCounts gathers raw usage activities and consolidated usage statistics reports respectively, for OpenAIRE Research Graph products and from the network of OpenAIRE content providers (repositories, e-journals, CRIS). This is realized 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 Code of Practice[1].

UsageCounts Service, allows sharing of usage statistics across the above distributed network and provides significant added value for different stakeholders. On the content 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. Moreover, OpenAIRE UsageCounts Service not only facilitates the above added-value services, but also allows the aggregation of usage data about research products which are published in several places.

Being aware of the sensitivity of this usage data, legal constraints are considered regarding the EU General Data Protection Regulation (GDPR). 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

[1] projectcounter.org

examine correlations with other types of metrics, e.g. bibliometric and webometric. This service is integrated with the repository dashboard, the OpenAIRE portal and API for 3rd party reuse.

Almost 200 Content providers from all over the globe are participating in the UsageCounts Service. The service has collected usage statistics information for 3.5 millions research products, manifested in 100 millions views and 380 millions downloads.

UsageCounts Architecture

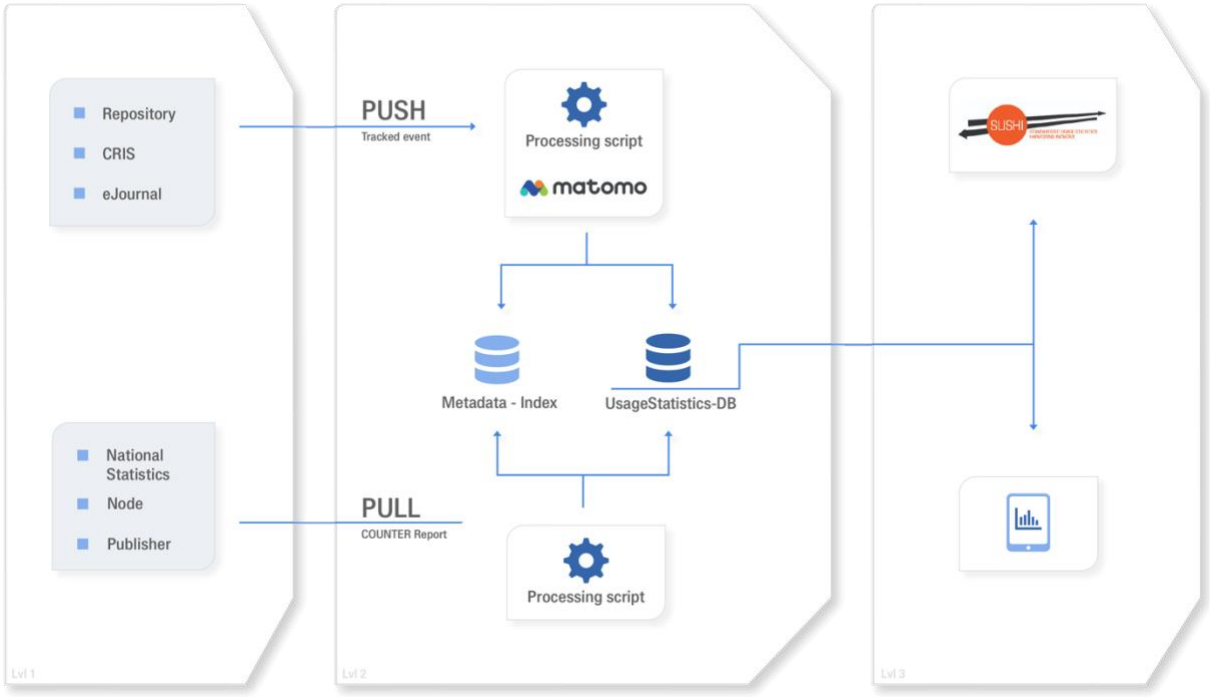


Figure 2. UsageCounts Service Architecture & Workflows

Figure 2 illustrates the two approaches exploited for the collection of usage data by UsageCounts, named *PUSH* and *PULL*. The PUSH approach, is the default workflow and employs the Matomo[2] analytics platform to track and store raw usage events. OpenAIRE Content providers might either deploy a tracking code in the form of DSpace plugins or EPrints patches, based on their platform type, or use a generic script that pushes usage activity using Matomo's HTTP API. Matomo's HTTP API can also be utilized by content providers (eg. LaReferencia) to develop their own applications and push raw usage activity to Matomo, given that the following parameters are included:

Parameter	Description
rec	Required for tracking, must be set to one

2

idsite	the ID of the content providers
idVisit	a visitor/session ID (an 8 byte binary string) automatically created by Matomo
cip	the IP address of the visitor (optionally anonymized)
action_name	the title of the item being accessed
url	the url of the requested item
download	the url of the item, in case of a download
timestamp	the date & time of the request, automatically created by Matomo
cvar	A custom variable to store the OAI-PMH Identifier, of the item being viewed/downloaded
ua	the Web Browser and the operating system of the visitor
urlref	The url linked to the item requested
token_auth	32 character authorization key used to authenticate the API request

Information is transferred offline to OpenAIRE's Statistics DBs for further processing and aggregation with information from OpenAIRE Research Graph.

A different approach for the UsageCounts service, named PULL is also depicted in Figure 2. Following this approach, OpenAIRE content providers or usage statistics aggregation services (e.g. IRUS-UK) offer a bulk download method for the usage data. In particular, PULL approach supports the gathering of consolidated statistics reports using protocols such as SUSHI-Lite. These statistics are also stored in OpenAIRE's Statistics DB and aggregated with information from OpenAIRE Research Graph.

Usage Statistics are finally deployed via OpenAIRE's Portals, like Explore, Provide, UsageCounts Portal, or can be retrieved by a Sushi-Lite API endpoint that complies to COUNTER Code of Practice R4.

Zenodo & Episciences Use Cases

Zenodo has already realized its integration with UsageCounts in the first months of the project. Zenodo offers download and view statistics that will be displayed alongside articles hosted on Zenodo. Usage activity, including both views and downloads, is pushed to OpenAIRE's Matomo

analytics platform via an application developed by Zenodo, using the parameters described in the previous section. Currently, Zenodo is working on making usage statistics for research data available through Datacite using the COUNTER Code Of Practice for Research Data[3], which would then be collected up by UsageCounts.

Episciences will join OpenAIRE UsageCounts service as a journal aggregator. A customized version of the generic tracker will be used, i.e. Using the PUSH approach, to send raw usage activity to OpenAIRE's Matomo analytics platform. This customization would allow the tracking of other Persistent Identifiers (PIDs) for journal items, if the oai-pmh identifier is not available.

Activity 3.3.2 episciences.org to integrate Zenodo.org APIs for deposition of pre-prints and vice versa, to enable submission of pre-prints to journals via Zenodo.og

Episciences model of interaction with repositories

The Episciences platform uses OAI-PMH protocol to interact with open repositories. The metadata is automatically retrieved with this protocol, i.e. pulled from the repository's endpoint. Additional information may be obtained using *ad hoc* API depending on the options offered by the repositories.

When documents are published, the platform usually communicates the information that the preprint has been reviewed, endorsed and published by a journal. For arXiv, Episciences provides an XML feed[4] parsed by the repository. arXiv will use it to update bibliographical references on their web pages. For HAL, an internal API is used to enrich the metadata on the HAL document's landing page.

Zenodo will be added as a source of preprints for Episciences journals. It will allow authors to submit preprints from Zenodo to a journal. The next step will be to allow the other way around: the submission of a preprint from a journal to Zenodo using its APIs. This feature is new from the Episciences point of view and for the moment will be limited to Zenodo.

Zenodo will also be added as a source to allow authors to link datasets and software to their publications, following a pattern already implemented by Episciences.org with SoftwareHeritage and HAL.

Figure 3 illustrates the data flows between episciences.org as an overlay on top of repositories and software archive. This activity, will be facilitated by activities in WP7, focusing on the identification of an interoperability framework for the deposition and fetching of research

[3] https://www.projectcounter.org/code-of-practice-rd-sections/foreword/
[4] https://arxiv.org/help/bib_feed

products into and from EOSC data sources. Episciences.org will therefore enable its data flows with any repository implanting the framework.

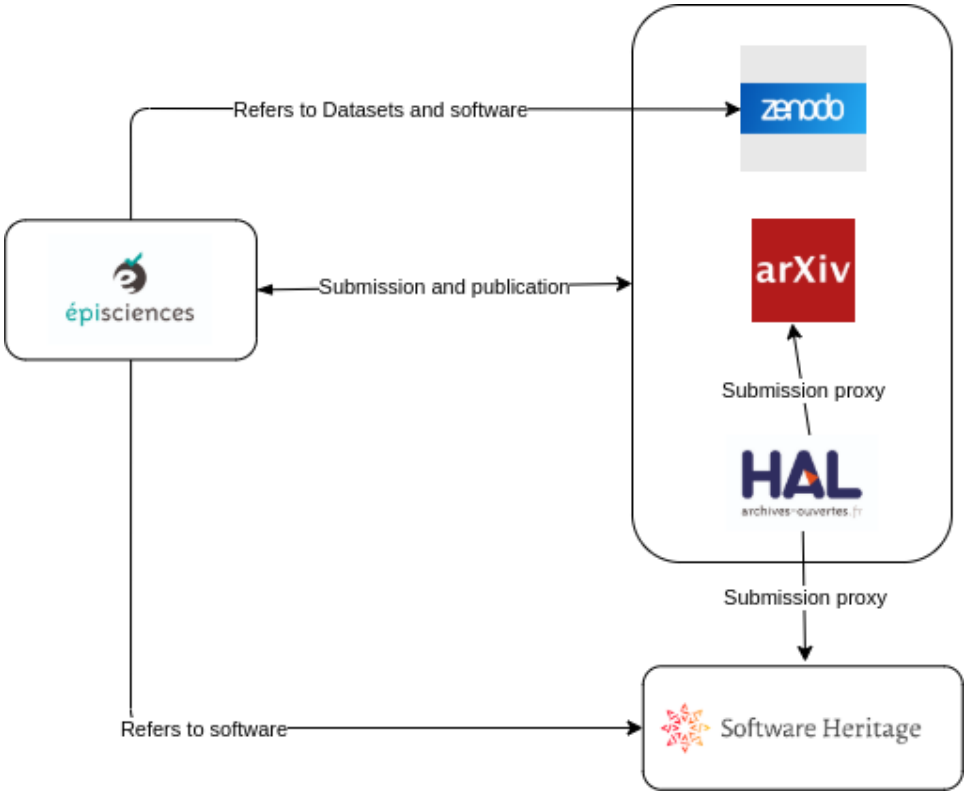


Figure 3. Interactions between episciences.org and services like Zenodo/HAL/Arxiv

Activity 3.3.3 episciences.org to integrate with OpenAIRE PROVIDE to receive Broker notifications

OpenAIRE PROVIDE (https://provide.openaire.eu) is a one-stop-shop web dashboard where content providers (repository, data archive, journal, aggregator, CRIS system) interact with OpenAIRE. It provides the front-end access to many of OpenAIRE's backend services: Registration, Validation, UsageCounts and Broker.

By registering in PROVIDE, a content provider joins the OpenAIRE network and contributes with its metadata records and full-texts to the OpenAIRE Research Graph. To join the OpenAIRE network, repositories and archives must export metadata records about scholarly objects of any type (published scientific literature, pre-prints, research data and software, scientific workflows, patents, and other types of research products) according to the OpenAIRE interoperability guidelines. If applicable, the content provider can give the consent to OpenAIRE to also collect the full-texts of Open Access publications. OpenAIRE will run its full-text and data mining algorithms to further enrich metadata records with links to projects, publications, datasets, software, organisations, research infrastructures and terms from standard classification

schemes. The aggregated metadata records will be included in the OpenAIRE Research Graph and merged with duplicates that OpenAIRE might have collected from other providers to form a richer record that is the union of the information available from each provider. The content provider can get the enrichments OpenAIRE introduced to its records by using the OpenAIRE Broker service available in the PROVIDE dashboard.

In particular, the Broker service identifies the following categories (called topics) of enrichments:

- Additional PIDs of publications (e.g. DOIs)
- Links to projects
- PIDs for authors of publications (i.e. ORCID)
- Links to Open Access versions
- Additional classification subjects (e.g. subjects from standard schemes like ACM, JEL and DDC)
- Abstracts identified in duplicate publications
- Missing publication dates

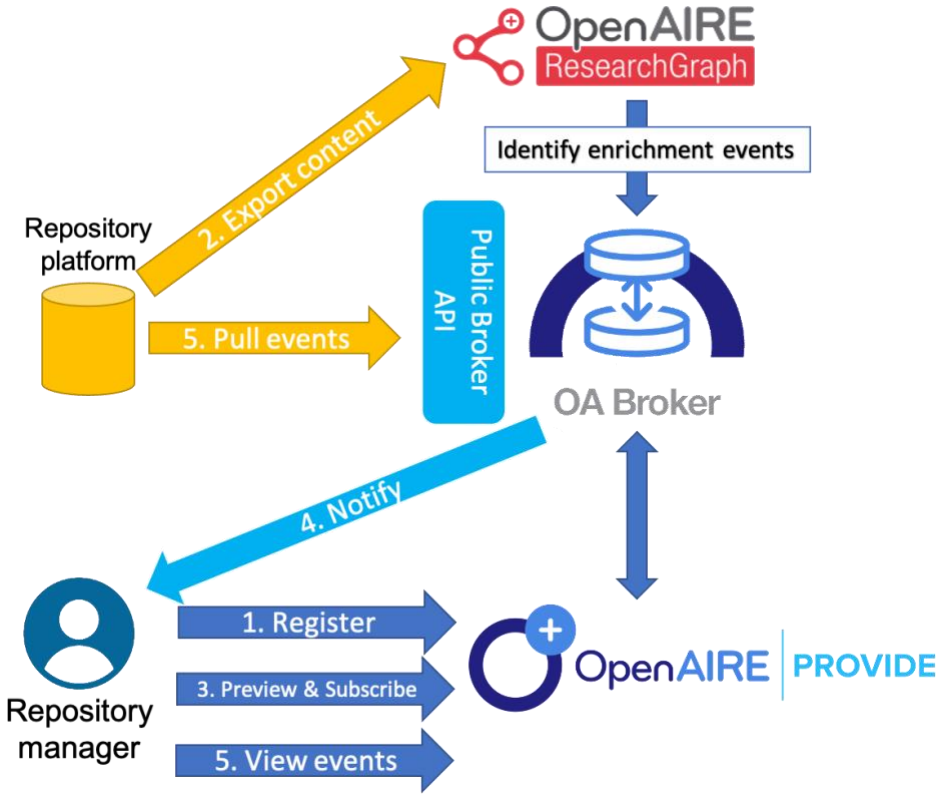


Figure 4. The OpenAIRE Broker and interactions with repositories

The PROVIDE dashboard offers the possibility to preview a set of enrichments, organized in categories, called *topics*, that represent the different types of enrichments OpenAIRE can

identify. For each topic the preview consists of 100 “enrichment events”, a subset of all the possible enrichments pertinent to a given repository in the OpenAIRE Graph, that the user can explore by applying filters on different criteria. The total number of events that can be potentially built is highlighted in the UI. Repository managers can create subscriptions for specific topics, which include the filtering criteria used to analyze the preview, or can subscribe to all the available topics with no restrictions at once. Once the repository manager creates a subscription, the algorithm analyzing the OpenAIRE Research Graph will produce the full set of enrichments for the repository. The enrichments will be made available as notifications in a dedicated section in the PROVIDE dashboard to be further checked as well as through the broker service API for programmatic access[5]. Repository managers are notified about the existence of new events matching their subscriptions by email.

episciences.org and the Broker service

Episciences.org will register to the PROVIDE Dashboard and export its records so they will be included in the OpenAIRE Research Graph. Episciences will update its OAI-PMH endpoint to comply with the OpenAIRE guidelines to provide information in a suitable format.

Using the PROVIDE dashboard, the manager of episciences.org will check the preview of enrichments events and will subscribe to the topics of interest. List of preliminary interesting topics:

- Additional PIDs of its publications;
- Links to projects;
- ORCID to be associated to publications authors;
- Additional classification subjects.

The notified enrichments events will be used to enrich the metadata records in the episciences.org system. Episciences.org will evaluate the most proper approach for integrating the enrichments (manual integration vs automated process using the Broker API) when the first set of events will be available.

Activity 3.3.4 episciences.org to integrate with Scholexplorer/OpenCitations API

Scholexplorer and OpenCitations services maintain two collections of aggregations of citation links between datasets-articles and articles-articles respectively. Scholexplorer links are collected from OpenAIRE scholarly communication sources, while links from OpenCitations from the publishers or sibling initiatives – currently, the main part of the citation data available are coming from Crossref, but additional sources will be added in the next years. Both services offer APIs to enable resolution of persistent identifiers (DOIs, PMCIDs, ArXiv, PDBs, etc.) to fetch the list of

[5] OpenAIRE Broker Service Public API Swagger documentation: http://api.openaire.eu/broker

related Scholix objects, i.e. relationships between the PID's object and the PIDs of the cited/citing objects, together with the relative semantics.

Several services worldwide are today embedding into their websites calls to the APIs of both services, in order to show contextual information about scientific articles they display. Similarly, episciences.org will implement widgets to access and show related citations to articles and datasets when these are available from the journals article pages.

3 OPENCITATIONS

OpenCitations is committed to one activity:

- Activity 3.3.5 OpenCitations to integrate citation links into the OpenAIRE Research Graph component of OpenAIRE MONITOR and vice versa.

Activity 3.3.5 OpenCitations to integrate citation links into the OpenAIRE Research Graph component of OpenAIRE MONITOR and vice versa.

OpenCitations is an Open Science infrastructure organization dedicated to the publication of open bibliographic and citation data by the use of Semantic Web (Linked Data) technologies. The main collection of data OpenCitations makes available is called OpenCitations Indexes. The Indexes contain information about the citations themselves, in which the citations, instead of being considered as simple links, are treated as first-class data entities in their own right. This permits each Index to endow each citation with descriptive properties, such as the date on which the citation was created, its timespan (i.e. the interval between the publication date of the cited entity and the publication date of the citing entity), and its type (e.g. whether or not it is a self-citation). As of 31 March 2021, OpenCitations Indexes contains more than 759 million citations between more than 60 million entities.

In addition to the OpenCitations Indexes and other smaller collections (i.e. the OpenCitations Corpus and the Open Biomedical Citations in Context Corpus), the OpenCitations ecosystem introduced in Figure Figure 5 includes:

- The OpenCitations Data Model (based on the SPAR Ontologies, http://www.sparontologies.net), which is the data model used to store all the data made available;
- All the software used for running OpenCitations services and for producing the data released with open source licenses and available in GitHub;
- A series of online services including REST APIs, SPARQL endpoints, dumps, and search and query interfaces.

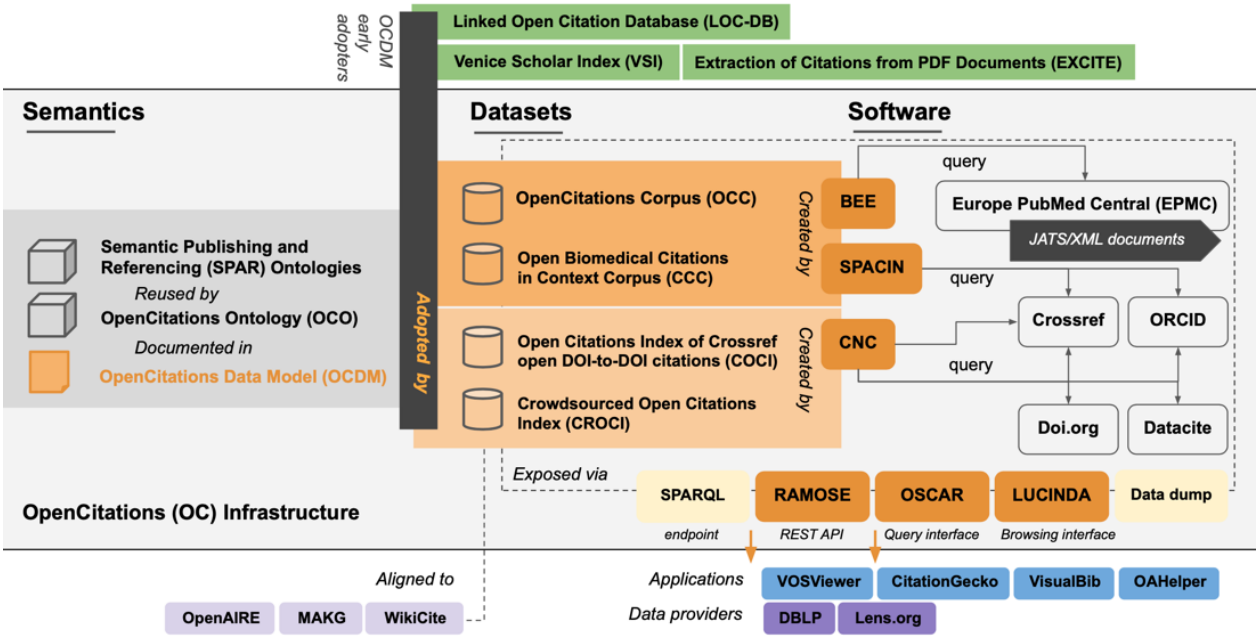


Figure 5. An overview of the OpenCitations ecosystem.

OpenCitations will become a precious source of information for the OpenAIRE Research Graph, via integration of the OpenCitations Scholix dump. This action will make sure all open citations between articles will be available in the same collection together with the citations between article and datasets of Scholexplorer.

On the other hand, OpenAIRE produces links between articles via full-text mining. Such links are CC0, as they are derived data from Open Access pre-prints, and will be made available to OpenCitations to potentially enrich its collection – e.g. by creating a new OpenCitations Index of OpenAIRE citations to be accessed by means of either its own Open API or in federation with the other OpenCitations Indexes.

The optimal scenario is one where OpenAIRE is a data source of citation links for OpenCitations and, vice versa, OpenCitations is the unique data source for citations for OpenAIRE.

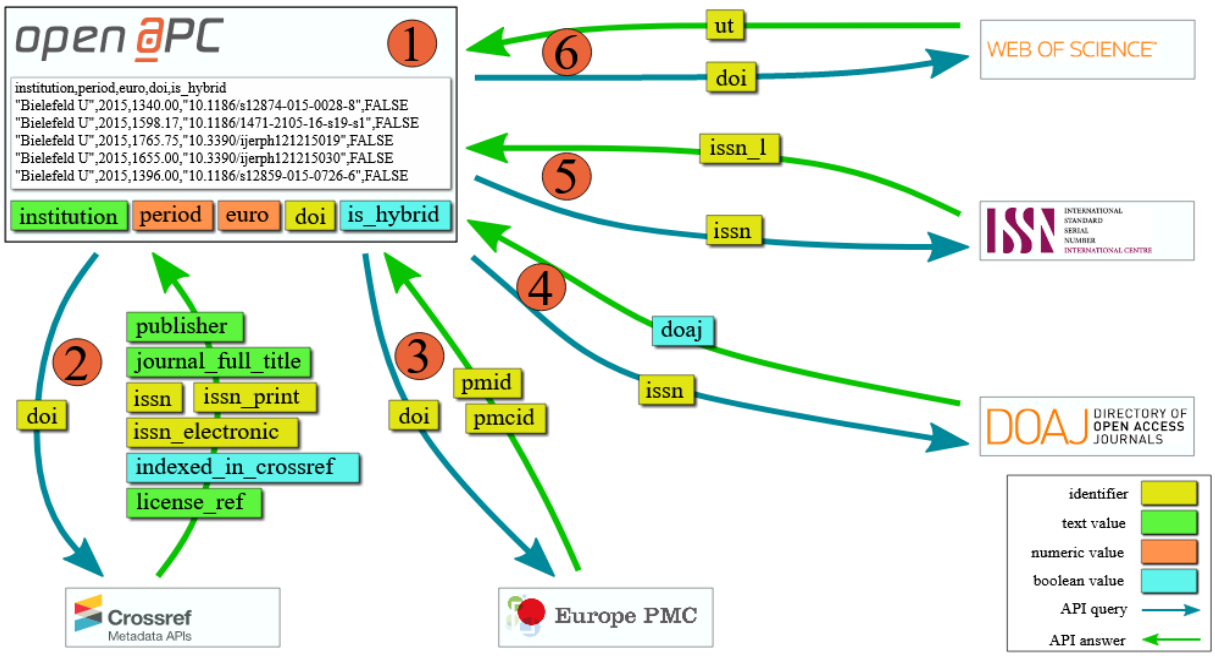
4 OpenAPC

OpenAPC is engaged in one activity:

- Activity 3.3.6 OpenAPC to integrate APC data into the OpenAIRE Research Graph component of OpenAIRE MONITOR to include extra information for monitoring purposes (institutional, OS monitor, etc.)

OpenAPC[6] collects, aggregates and publishes APC and other cost data from participating institutions. It aims at transparency, comparability and tracking of cost developments in the field of OA publishing. OpenAPC allows libraries, funding agencies, researchers, developers and 3rd party services to keep track and provide access to the Open Access record of European expenditure for APC or other cost data, e.g. from transformative agreements, across publishers, journals, academic institutions and countries. All OpenAPC data is made freely available under the Open Database License (ODBL). OpenAPC complies with current recommendations for cost transparency in an OA based scholarly publication system. Major Plan S cOAlitionS members and supporters e.g. Wellcome Trust, FWF, or the Bill & Melinda Gates Foundation are already contributing data to OpenAPC.

The OpenAPC dataset collects a very small set of attributes from the institution/organization that participates. These attributes are namely: *institution name*, *period of apc*, *euro*, *doi*, and whether or not the paper is published in a *hybrid journal*. This fundamental information will be enriched on different workflow steps. The workflow steps and processing are shown in the figure below.



The workflow steps are:

1. Collecting information from institutions or organizations
2. Enriching via CrossRef

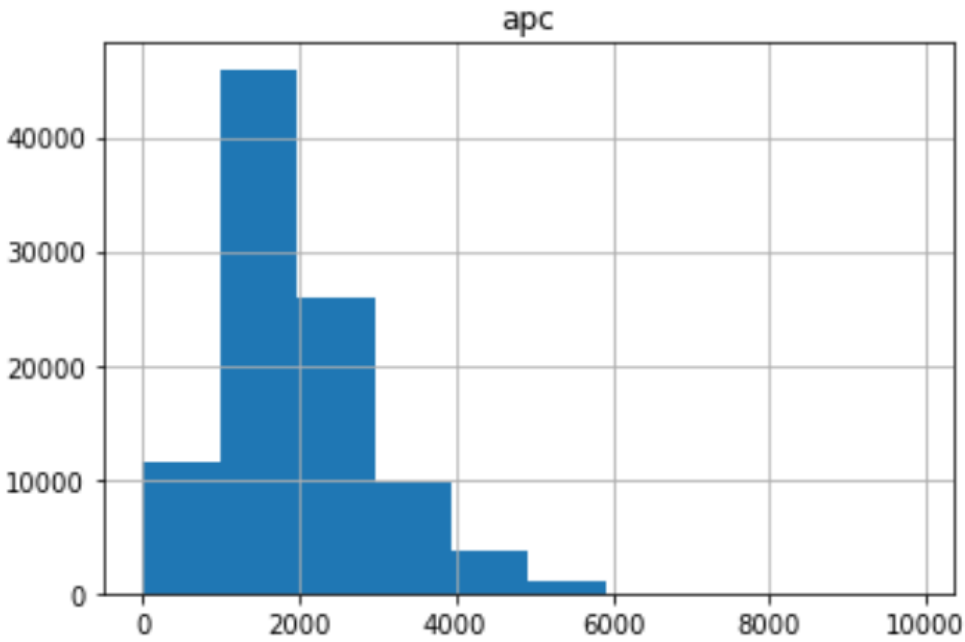
[6] https://openapc.net/

3. Enriching via Europe PubMedCentral
4. Enriching via Directory of Open Access Journals (DOAJ)
5. Enriching via Internal Standard Serial Number (ISSN)
6. Enriching via Web of Science (WoS)

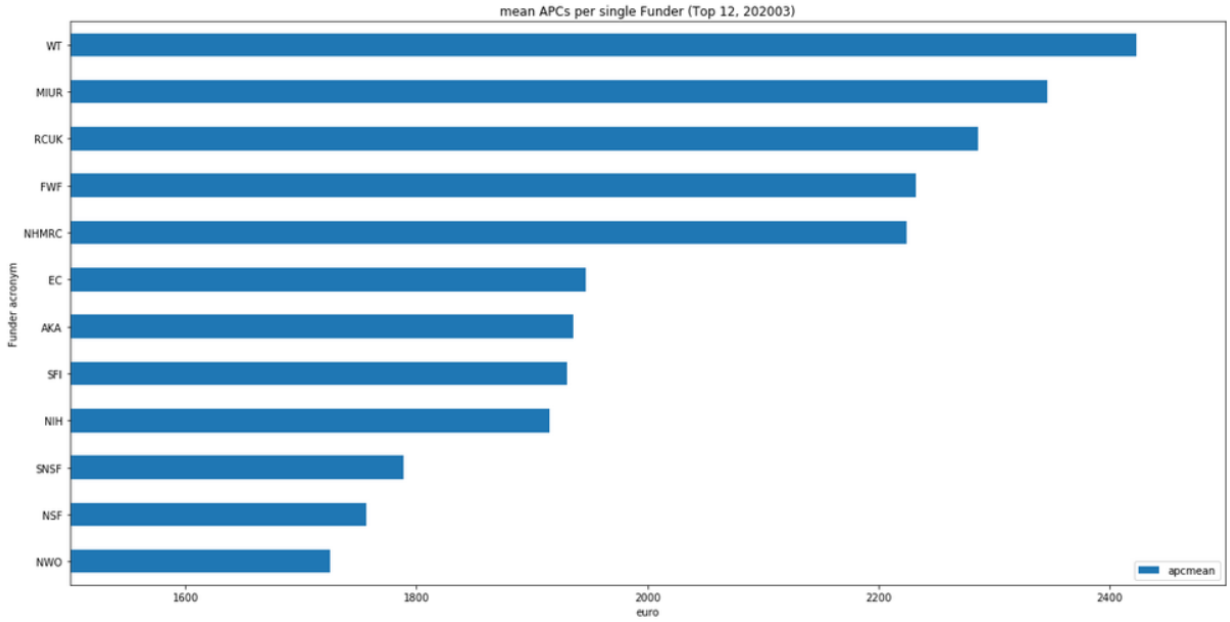
That enrichment process serves to complete information for further evaluation in the Online Analytical Processing service (OLAP). OLAP already provides a REST-API for direct integration into OpenAIRE services.

The APC dataset enriched in this way, in combination with the OpenAIRE Research Graph, offers the possibility of mapping cost data on the projects and funders linked to the publications. The final analyses of the two datasets during the OpenAIRE-Advance project gave us an indication of the many possibilities for analysis.

The comparison in March 2020 of OpenAPC dataset and the OpenAIRE Research Graph revealed an overlap between 97-98%.



The APC average of the overlapped records was around 1,979 EUR with a total sum of more than 190 million EUR. The figure below shows the funding of APCs costs by a single funder.



With the continuous integration of the APC dataset into the graph, there are opportunities for ongoing analysis, like shown above. Secondly, the OpenAPC initiative has also started to collect cost data on Book Processing Charges (BPC) on Open Access Books and Monographs and Open Access Transformative Agreements.

Actions

- Attracting more institutions and organisations to participate in the OpenAPC initiative
- Onboarding of OpenAPC into OpenAIRE and EOSC catalogue
- Continuous collecting of KPI statistics
- Refinement and enhancement of enrichment workflows
- Preparation, collection and integration of cost data on OA books and monographs, and on transformative agreements