graspos
open research assessment dataspace

DOI 10.5281/zenodo.8302198

Infrastructure Architecture

Deliverable D4.1

Due Date of Deliverable	31 August 2023 (M8)
Actual Submission Date	31 August 2023
Work Package	WP4
Tasks	T4.1
Type	R - Report
Approval Status	Submitted
Version	1.0
Number of Pages	27

The information in this document reflects only the author's views and the European Commission is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/her sole risk and liability.

Abstract

This report presents the high-level architecture of the GraspOS federated infrastructure, which aims to support responsible research assessment processes and enable the creation of a Research Assessment Dataspace. More specifically, we first elaborate on the motivation and the functional goals, and then we describe in more detail its most important components. This is the first version of this report, and improved versions will be published later on during the lifetime of the project.

This project has received funding from the European Union's Horizon Europe framework programme under grant agreement No. 101095129. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.

Revision history

VERSION	DATE	REASON	REVISED BY
0.0	21/04/2023	First draft created	Andrea Mannocci
0.1	03/08/2023	Agreement on structure & References	Thanasis Vergoulis
0.2	10/08/2023	Intermediate version	Giulia Malaguarnera
0.3	13/08/2023 and 24/08/2023	Peer review	Ludo Waltman and Laura Himanen
0.4	29/08/2023	Peer review comments addressed	Giulia Malaguarnera and Thanasis Vergoulis
1.0	30/08/2023	Final Version after proofreading	Giulia Malaguarnera

Author List

ORGANISATION	NAME	CONTACT INFORMATION
CNR-ISTI	Andrea Mannocci	andrea.mannocci@isti.cnr.it
ARC	Thanasis Vergoulis	vergoulis@athenarc.gr
ARC	Serafeim Chatzopoulos	schatz@athenarc.gr
UNIBO	Silvio Peroni	silvio.peroni@unibo.it
OpenAIRE	Giulia Malaguarnera	giulia.malaguarnera@openaire.eu
OPERAS	Suzanne Dumouchel	suzanne.dumouchel@operas-eu.org
CWTS	Clifford Tatum	c.c.tatum@cwts.leidenuniv.nl

Table of Contents

1. Executive summary 6
2. Introduction 7
3. Functional goals 8
4. Architectural overview 10
4.1. GraspOS Data Asset Sources 12
BIP! DB 13
OpenAIRE Graph 13
OpenCitations data 14
OPERAS metrics 15
ScholExplorer 15
Usage Counts 16
EOSC Accounting for Services 16
Assessment portfolios 17
Open Science Assessment Registry (OSAR) 18
4.2. Interoperability & access layer 18
4.3. Enrichment services 20
BIP! Citation Classifier 20
BIP! Services - Ranker 20
EC KIP OS indicators 21
OpenAIRE Broker 21
OpenAIRE IIS text-mining modules 21
OpenAIRE metadata validator 22
SCRE pipeline 22
Semantic citation classifier 22
4.4. Monitoring services 23
BIP! Scholar 23
EOSC accounting for services 24
OpenAIRE institutional monitor dashboard 24
OpenAIRE Researcher Profile 24
Open Science Observatory 24
EOSC Observatory 25
4.5. Data registry 25
5. Conclusions 26
6. References 27

Abbreviation List

- EOSC - European Open Science Cloud
- IIS - Information Inference Service
- ORAD - Open Research Assessment Dataspace
- OS - Open Science
- OSAF - Open Science Assessment Framework
- OSAR - Open Science Assessment Registry
- RDA - Research Data Alliance
- RFO - Research Funding Organisations
- RPO - Research Performing Organisations
- RRA - Responsible Research Assessment
- SKG - Scientific Knowledge Graph

1. Executive summary

Research assessment covers a pivotal role in a variety of applications in research, ranging from informing the hiring and promotion of research staff to shaping future strategic investments and policymaking.

However, widely used assessment practices face several challenges that cannot be underestimated, such as the use and abuse of a small pool of often misunderstood and non-transparent quantitative indicators, or the reliance on evidence from proprietary data sources with restricted access, to name a few.

To address the challenges of *Open-Science-aware Responsible Research Assessment* (OS-aware RRA), this deliverable introduces the GraspOS Federated Infrastructure aiming at enabling the creation of a Research Assessment Dataspace that is transparent, inclusive, and interoperable. The GraspOS Federated Infrastructure enables the integration of diverse data sources and makes them interoperable. This will allow for improved coverage of research products (input and output from research projects) and more comprehensive assessments that are based on a wider range of evidence.

The infrastructure is designed to be modular and scalable, allowing for the integration of new components and data sources as needed. It is also designed to be open and accessible, enabling the research community to contribute to its development and use. The infrastructure is built on a federated architecture, which means that it is distributed across multiple components. This enables the infrastructure to be more resilient and adaptable to changing needs and contexts.

Overall, the GraspOS infrastructure is composed of a set of federated Data Asset Sources providing access to and delivering metadata for research products, pre-calculated research performance indicators, as well as unstructured data such as scientific texts, narrative CVs and data on Open Science (OS) practices. Such sources are flanked with a set of federated monitoring and enrichment services, which can be used to handle the transformation, manipulation, and analysis of data within the infrastructure, present and visualise research assessment evidence and indicators at different levels of granularity (e.g., researcher, organisation, country) to support OS-aware RRA and to monitor the uptake and impact of OS. A Data Registry documenting the federated resources complements the infrastructure.

2. Introduction

Research assessment refers to the process of evaluating research activities, outputs, practices, and roles in terms of various characteristics such as their quality, novelty, scientific/financial/societal impact, usage, and accessibility. This includes the process of supporting researchers in their career progression, evaluating research performed by Research Performing Organisations (RPOs) and measuring the scientific, societal, and economic impact of research projects/grants for Research Funding Organisations (RFOs). Consequently, research assessment plays a crucial role, ranging from informing the hiring and promotion processes in RPOs to shaping strategic investments and policymaking.

Commonly used research assessment practices face major challenges. In particular, there is a growing recognition that assessing research and researchers needs to be done in a more responsible way. This requires a shift away from over-relying on a small pool of quantitative indicators, which are usually not calculated transparently and often misunderstood, towards a more comprehensive approach that considers different aspects and merits of research work and incorporates qualitative evidence. In addition, since research assessment practices shape research practices, they have to be well-thought-out to avoid amplifying problems such as the "Publish-or-Perish" culture and the Matthew effect (Merton, 1968), or fostering bad research practices (Strinzel, 2021). In response to these challenges, various initiatives for Responsible Research Assessment (RRA) have been launched (Moher D, 2020).[1,2,3]

Furthermore, research assessment processes should be tailored to the scope and the context of the evaluation. For instance, the research field, the purpose of the assessment process, the mission or the role of the evaluation subject, and even the country in which the assessment is performed may all represent unique characteristics that should be taken into consideration. To account for aspects like these, the INORMS Research Evaluation Group[4] has designed the SCOPE Framework, which is a practical guide for planning and conducting a research assessment responsibly following five steps: start with what you value, context considerations, options for evaluating, probe deeply, and evaluate your evaluation (SCOPE, 2023).

Another important problem is that research assessment practices currently often rely on indicators and evidence from proprietary data sources behind paywalls, hindering the transparency of the assessment processes and restricting the ability of the research

[1] The Declaration of Research Assessment (DORA), https://sfdora.org
[2] Leiden Manifesto for Research Metrics, http://www.leidenmanifesto.org
[3] Coalition for Advancing Research Assessment (CoARA), https://coara.eu
[4] INORMS, https://inorms.net

community to independently scrutinise the findings or to offer alternative interpretations of the data. To make matters worse, these data sources are not designed to be interoperable; hence, it is difficult to combine their contents to achieve improved coverage and capitalise on their plurality of views of the global research track record. Furthermore, the evaluation outcomes are deeply bound to the data source selected for the assessment process, as moving to another data source could, in general, yield profoundly different results.

Finally, research assessment practices often focus on traditional research outputs, mainly on English-language journal articles and their impact in terms of citations, and fail to properly acknowledge efforts to make research processes more transparent and research outputs more open. Although Open Science (OS) initiatives have gained popularity in recent years, reforming research assessment to acknowledge the contributions researchers and institutions have in OS is essential. A movement towards enabling an *Open-Science-aware Responsible Research Assessment* (OS-aware RRA) is important to encourage and reward the adoption of OS practices. At the same time, organising research assessment on the basis of OS principles, through the use of open data sources and transparent indicators, is increasingly seen as a precondition for realising the ambition of making research assessment more responsible. This shows that OS and RRA need to mutually reinforce each other.

GraspOS aims to alleviate the aforementioned problems by designing and delivering an *Open and Federated Research Assessment Infrastructure* (formerly known as “Federated Open Metrics Infrastructure - FOMI”). This infrastructure paves the way for the creation of an *Open Research Assessment Dataspace* (ORAD) by aggregating open resources (e.g., next-generation metrics and indicators, data, tools, services, and guidance offered by different sources) that can catalyse the implementation of policy reforms towards an OS-aware RRA framework.[5] The value of this infrastructure will be evaluated and showcased in practice by the GraspOS piloting activities (WP5) at different levels of granularity: from a researcher (individual/group) and institutional to an organisational and country level.

In the following sections, we offer a high-level description of the architecture to support the aforementioned infrastructure (and the Research Assessment Dataspace). It should be noted that this is the first version of the GraspOS architecture; updated versions of this report will be released at later points of the project life span.

3. Functional goals

[5] This framework (also known as Open Science Assessment Framework - OSAF) will be developed based on the recommendations made by WP2.

The GraspOS architecture should support and facilitate the following functional goals:

- The GraspOS infrastructure should federate multiple open data sources that can offer data assets valuable for the implementation of OS-aware RRA processes, paving the way for the creation of an Open Research Assessment Dataspace (ORAD). This dataspace will increase transparency in research assessment processes, helping to move away from paywalled data sources. The GraspOS architecture will support the implementation of this dataspace, supporting important functionalities such as communication and coordination between data providers and consumers, data discovery, and data integration.
- Additionally, the infrastructure should support the promotion of OS and collaboration, and the development of new research assessment indicators and methods.
- All GraspOS data sources and services should be interoperable by design, being compliant with relevant interoperability specifications and guidelines. The focus should be on the design of interoperability specifications and guidelines that will facilitate the federation of scholarly data sources and enable access to the respective data assets by end-users and added-value services. These specifications and guidelines will essentially create a set of "onboarding" requirements that can be used for the federation of additional sources in the future. The specifications will be compliant to relevant EOSC guidelines and well-established relevant standards.
- Regarding the contents of the federated data sources, the aim will be for them to collectively offer various types of data assets that are useful for the implementation of OS-aware RRA processes. Indicatively, the objective is to offer bibliographic records, relationships between research-related entities (e.g., authorship linkages, citations) and information for the relevant semantics (e.g., author contribution roles, citation intent), pre-calculated research performance indicators, usage data for research products (e.g., views and downloads of publications), scientific texts (e.g., contents of Open Access publications), research service usage indicators and accounting metadata (e.g., from EOSC accounting system), narrative CVs, research assessment protocols, and, of course, data on practising OS.
- Regarding the added-value services that we are going to build upon the above-mentioned data sources, the focus will be on two categories:
  - *Enrichment services* to enrich research product metadata records with additional metadata (e.g., missing attribute values, link semantics, impact metrics) and missing links (e.g., missing citations, affiliation/authorship links).

  - *Monitoring services* to report and visualise information and indicators to support OS-aware RRA and/or to monitor the uptake and impact of OS from multiple perspectives (e.g., institutional-level, national-level, scientific, and societal).
- Most of the services will be based on the extension and adaptation of relevant, well-established services already provided by the technology-providing partners in the GraspOS project (e.g., OpenAIRE, UniBo, OPERAS, ATHENA). The developments will partly depend on the requirements of the Open Science Assessment Framework (OSAF; WP2) to ensure that the GraspOS services will facilitate a systematic approach to the implementation of OS-aware RRA processes. In addition, special requirements coming from the use cases of GraspOS pilots (WP5) will be taken into consideration as well.
- Finally, GraspOS services will be integrated into EOSC (e.g., being available in the EOSC marketplace), contributing to the expansion of EOSC with important services in the field of research assessment and OS monitoring. In addition, all services will leverage existing EOSC components where applicable to reduce development efforts and increase the level of integration with the EOSC ecosystem.

4. Architectural overview

Figure 1 offers a high-level representation of the architecture of the GraspOS Open and Federated Research Assessment Infrastructure. Its core consists of the following:

- A set of ***Data Asset Sources***, which provide both the raw and the structured data that can be used to directly support OS-aware RRA processes or to calculate indicators or collect evidence that can be of value in these processes.
- An ***Interoperability & Access Layer***, which facilitates the integration and harmonisation of (meta)data from various sources and formats within the infrastructure and provides recommendations on data models and API specifications that facilitate accessing the scholarly data sources by users and added-value services.
- A set of ***Enrichment Services***, which can be used to handle the transformation, manipulation, and analysis of data within the infrastructure to produce enrichments that can be of value in the context of OS-aware RRA.
- A set of ***Monitoring Services***, which focus on reporting and visualising *i)* research assessment evidence and indicators at different levels (e.g., researcher, organisation, country) to support OS-aware RRA processes, and *ii)* to monitor the uptake and impact

of OS from multiple perspectives (e.g., institutional-level, national-level, scientific societal).

- A ***Data Registry***, which allows users to be informed about and understand the data included in the infrastructure. It serves as an inventory of the data sources and assets.

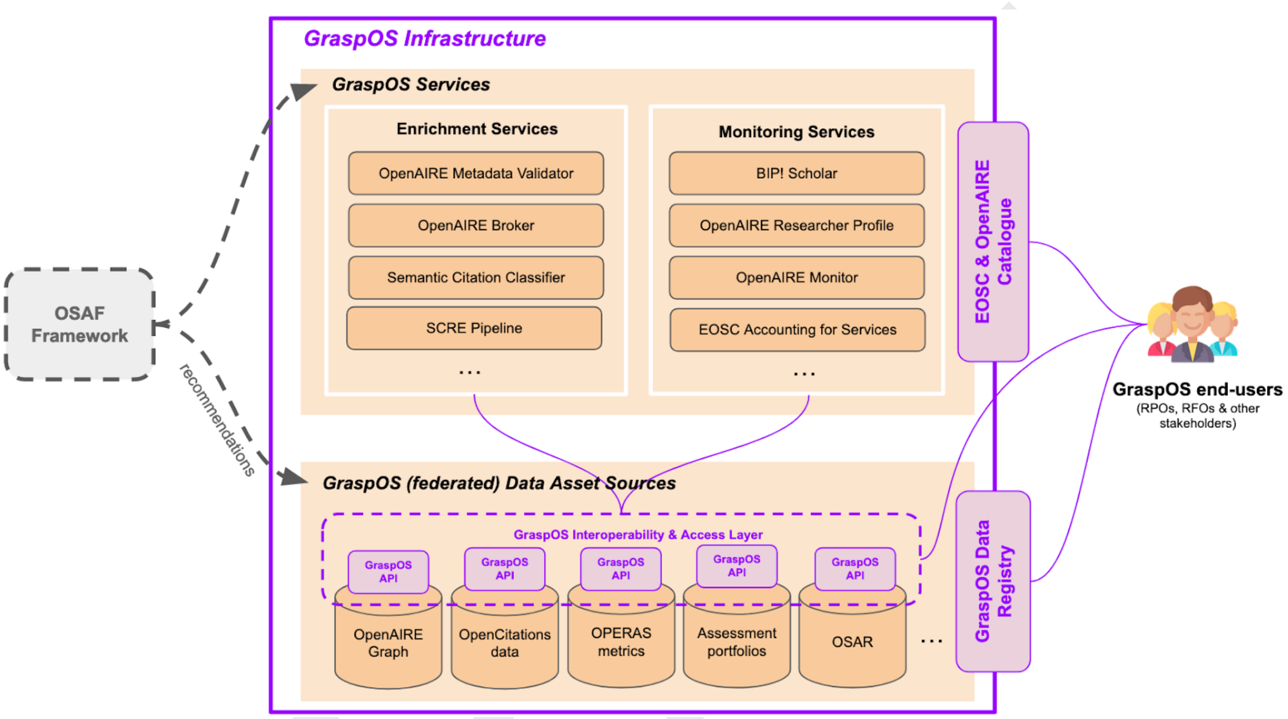

Figure 1 - The GraspOS infrastructure architecture

All federated data asset sources will be openly available and easily accessible through GraspOS-compatible APIs (more details in Section 4.2). In addition, all GraspOS services (enrichment and monitoring) will be onboarded in the EOSC Catalogue, either directly or by creating entries in the OpenAIRE Catalogue.[6]

All GraspOS data asset sources and service providers will follow the OSAF framework (WP2) recommendations to better support OS-aware RRA processes. In addition, requirements from the various GraspOS pilots (WP5) will be considered to facilitate the respective use cases.

In the following sections, we further elaborate on each of the components.

[6] OpenAIRE catalogue, https://catalogue.openaire.eu

4.1. GraspOS Data Asset Sources

GraspOS architecture aims to federate sources of open scholarly data assets that can be used to support OS-aware research assessment processes. These sources represent the data providers of the respective Research Assessment Dataspace, and their federation will be possible thanks to the common interoperability specifications for APIs and data models that the GraspOS project will adopt and extend (considering the outputs of relevant initiatives[7]). The federated data sources should be compliant with these specifications. This "onboarding" requirement will facilitate the consumption of their assets by the GraspOS services (see also Sections 4.3 and 4.4) and by research assessment processes directly. Moreover, it will standardise the process for the addition of extra sources in the future. More details on the respective interoperability specifications can be found in Section 4.2.

The federated sources will provide data assets of various types, including (among others):

- *Metadata records* covering important attributes of research-related entities (e.g., research products, researchers, organisations) and the relationships among them (e.g., authorship linkages, citations). This type of data can be used by users/consumers to calculate indicators or construct supporting evidence that can be valuable for OS-aware RRA processes. Special care will be given to information that can be conducive to RRA practices, such as authors' contribution roles and citation intents.
- A variety of *pre-calculated research performance indicators* that can assist OS-aware RRA processes. Specifically, we aim to provide an extensive set of indicators for research products, researchers, and organisations, including traditional measures (e.g., citation counts) and flanking them with novel indicators developed or extended within GraspOS (e.g., indicators included in BIP! DB) and in the context of other projects (e.g., PathOS[8]). This will include indicators for the usage of research products (e.g., views/downloads of publications) and research-related services (e.g., exploiting accounting and metrics services of the EOSC Core to build the layer that integrates usage data from research services). By combining traditional indicators with new ones and other relevant information, we aim to offer a more inclusive research assessment approach where different factors can be considered according to the assessment protocol at hand.

[7] Indicative examples are the outputs of the RDA Working Group on Scientific Knowledge Graphs - Interoperability Framework (SKG-IF) and the activities related to the implementation of the Metadata Schema and Crosswalk Registry (MSCR) from the FAIRCORE4EOSC project.

[8] PathOS project: https://pathos-project.eu

- *Narrative CVs and scientific texts* coming from Open Access publications, that can be used to support research assessment processes, reveal latent information for the assessment subjects, or/and facilitate the calculation of new-generation indicators based on extracted data.
- *Data on practising OS*, i.e., data about the participation/contribution of researchers to miscellaneous OS-related activities and their adoption of OS practices (e.g., open access publishing, self-archiving, data sharing, DMPs publishing).

GraspOS users/consumers (researchers and other stakeholders) will be able to explore the available federated sources through the GraspOS data registry and have programmatic access to them through appropriate APIs.

In the next paragraphs, we briefly describe the core data asset sources that will be federated into the GraspOS architecture. Please note that these are important indicative sources, yet not the only ones that will be used in the context of this project.

BIP! DB

BIP! DB is an open dataset[9] that provides citation-based impact indicators, offering insights into various dimensions of scientific impact. These indicators encompass citation counts, popularity (current impact calculated based on citation network analysis), influence (overall impact based on citation network analysis), and impulse (initial momentum in terms of citations received the first years after publication), providing a multifaceted view of a research product's scientific impact. To compute those indicators, BIP! DB collects citation data from Crossref, OpenCitations, and the OpenAIRE Graph to build a comprehensive citation network comprising over 130 million research products and more than 1.6 billion citation relations among them. The dataset is available in CC-BY.

OpenAIRE Graph

The OpenAIRE Graph[10] (formerly known as the OpenAIRE Research Graph) is one of the largest massive open scholarly record collections worldwide, consisting of metadata and links between research products, organisations, funders, funding streams, grants/projects, communities, and data sources. It is a public and transparent resource that aims to facilitate the discovery, monitoring, and assessment of science. The OpenAIRE Graph aggregates millions of metadata records from several trusted sources such as Crossref, Unpaywall,

[9] BIP! DB: https://doi.org/10.5281/zenodo.4386934
[10] https://graph.openaire.eu

ORCID, Microsoft Academic Graph, DataCite, as well as repositories registered, among others, in OpenDOAR, re3data.org, FAIRSharing.org, and the EOSC Service Catalogue.

OpenAIRE's methodological approach[11] emphasises operational quality criteria[12], and begins with openness and transparency, clearly presenting the underlying assumptions. To ensure comprehensive coverage and accuracy, it exploits multiple data sources from the OpenAIRE graph. Clarity and replicability are addressed through a detailed description of the construction methodology, enabling verification and ongoing updates by the scholarly communication community. Readiness and timeliness are achieved by leveraging established open databases and tested knowledge extraction technologies. Trust and robustness are emphasised through alignment with other assessment methods, facilitating operationalisation alongside them.

OpenCitations data

OpenCitations is an independent, community-led, and not-for-profit OS infrastructure organisation that publishes open bibliographic and citation data by using Linked Open Data (LOD) using Semantic Web technologies. OpenCitations provides two main collections. The first one is the OpenCitations Index[13], which contains more than 1.4 billion citation links between entities gathered from different sources, which include Crossref, DataCite, the National Institute of Health Open Citation Collection, and the OpenAIRE Graph. The other collection is OpenCitations Meta[14], which currently includes basic bibliographic metadata (title, date of publication, venue, identifiers, authors, publisher, etc.) of the entities involved in the citations included in the OpenCitations Index. The OpenCitations Index and OpenCitations Meta constitute valuable resources for bibliometric analyses, such as estimating the impact of research products and enabling their reproducibility. All OpenCitations data are licensed using a CC0 waiver[15], are downloadable in full (https://opencitations.net/download), and can be accessed programmatically through its querying services (https://opencitations.net/querying), which include several REST APIs.

[11] OpenAIRE Monitor documentation, https://monitor.openaire.eu/methodology/terminology#entities

[12] European Commission, Directorate-General for Research and Innovation, Monitoring the open access policy of Horizon 2020: final report, Publications Office, 2021, https://data.europa.eu/doi/10.2777/268348

[13] OpenCitations Index, https://opencitations.net/index

[14] OpenCitations Meta, https://opencitations.net/meta

[15] CC0 waiver, https://creativecommons.org/publicdomain/zero/1.0/legalcode

OPERAS metrics

OPERAS Metrics[16] is a usage and altmetrics platform for Open Access publishers in the Humanities and Social Sciences. It collects usage and impact metrics related to published Open Access content from many different sources (monographs, journals, repositories) and allows their access, display and analysis from a single access point. Metrics are displayed not only for the publisher's website but are also aggregated with those of other sites where a book is known to be available at. The OPERAS Metrics service provides an Open Source tool able to collect usage and alternative metrics for Open Access publications. The goal is to collect usage and impact metrics related to published Open Access content from many different sources and allow for their access, display and analysis from a single access point.

Most platforms are limited to collecting their own usage, while many do not even comprehend the complexity of the matter. The HIRMEOS project posed a groundbreaking approach that enables metrics collection and aggregation from third-party platforms, which is currently a manual job for many scholarly publishers that lack the funds to find a technical solution to the problem. While many argue that third-party metrics should not be collected until platforms achieve a uniform collection and reporting mechanism, OPERAS acknowledges the need for authors and publishers to have access to data – and that a unique collection mechanism must not be imposed upon the community. For this reason, a simple standard has been designed that allows easy cross-platform analysis through transparent tagging of data to its definition.

ScholExplorer

This dataset[17] contains Scholix links[18] exposed by the OpenAIRE ScholeXplorer service. It consists of 417+Mi bidirectional links (i.e. 975+Mi directed links) between literature-to-dataset and dataset-to-dataset involving 24+ Mi literature objects and 37+ Mi datasets. Links are collected from publishers (CrossRef, EventData), data centres (e.g., DataCite), institutional/thematic repositories, and life science databases (EMBL-EBI), and inferred by OpenAIRE via text-mining around 14Mi publication's PDFs. The dataset is structured in 30 compressed files, each of at most ~10 GB, for a total of ~328 GB. ScholExplorer data are also available in the OpenAIRE Graph and are, hence, available in the respective public dumps and through the OpenAIRE Graph API.

[16] OPERAS metrics service, https://metrics.operas-eu.org
[17] ScholExplorer dataset, https://doi.org/10.5281/zenodo.1200252
[18] Scholix, http://www.scholix.org

Usage Counts

Usage Counts Service[19] gathers usage events and consolidated usage statistics reports, respectively, from a distributed network of data providers (repositories, e-journals, CRISs) by utilising 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.[20] The Usage Counts Service, allows the sharing of statistics across the above-distributed network and provides 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.

Two different approaches are deployed to collect usage events. The 1st approach is called PUSH, where a datasource registers in the Service via the OpenAIRE's Provider Dashboard. Then, on the Server side, anonymised, real-time tracking is offered to collect real-time usage events using Plugins (DSpace), and patches (Eprints). The 2nd approach, named PULL, employs an offline workflow based on Sushi-Lite API[21], to transfer and store usage statistics reports from aggregators or other data sources. The usage Statistics built from information from these two approaches are published for consumption in portals, like OpenAIRE's explore, or machines via an API based on Sushi-Lite. Usage Counts data are incorporated in the OpenAIRE Graph and are, hence, available in the respective public dumps and through the OpenAIRE Graph API.

EOSC Accounting for Services

EOSC Accounting for Services[22] is a platform designed to efficiently collect, aggregate, and exchange metrics across various infrastructures, providers, and projects. The system provides a REST API, which accepts input from diverse resources, stores it in a database, and aggregates the incoming data. It also offers an intuitive user interface that allows clients to interact with the platform and access accounting data for specific time periods. All API resources are only accessible to authenticated clients, ensuring secure access to sensitive data.

[19] UsageCounts, https://usagecounts.openaire.eu

[20] https://www.projectcounter.org/code-of-practice-five-zero-two

[21] https://app.swaggerhub.com/apis/COUNTER/counter-sushi_5_0_api/5.0.3

[22] https://accounting.eosc-portal.eu and https://argoeu.github.io/argo-accounting

The key functionalities offered by the EOSC Accounting Service are

- Efficient collection, aggregation, and exchange of metrics.
- REST API that accepts input from diverse resources.
- Database storage and aggregation of incoming data.
- Intuitive user interface for accessing accounting data.
- Secure access to sensitive data through authenticated clients.

The Accounting Service is a system that provides a framework for organising and managing accounting data for a specific project, provider or installation. It involves various roles such as Project Admin, Provider Admin, and Installation Admin, each with a specific set of responsibilities.

One of the key elements of the Accounting Service is Metrics, which are quantitative measures used to assess and track the performance or usage of a service. A Metric Definition is a way of representing and describing the type of metric.

In the Accounting Service, Metric and Unit Type are essential components that allow for the collection and tracking of various types of Metrics. These types are included as part of the Metric Definition and provide greater flexibility and specificity in the Metrics that the Accounting System can collect. A Metric Type defines how physical quantities are collected over time, while a Unit Type expresses and measures physical quantities used in various infrastructures, service providers, and projects. Together, these elements enable users to gather and analyse data at different levels of granularity and with different units of measurement.

Assessment portfolios

Assessment portfolios facilitate the collection of inputs for research assessment, serving both as an account of the agreed evidence for a given assessment event and as a shared resource for conducting the assessment. Assessment portfolios are based on the Research Activity Identifier (RAiD), which has two components; a persistent identifier and metadata record for documenting the evidentiary and descriptive contents used in an assessment event. Templates are provided for the following assessment levels of aggregation: individual, research group, research institutions, research community, and country.

Open Science Assessment Registry (OSAR)

Also based on the RAiD, the assessment registry facilitates the publication of assessment protocols after the completion of an assessment event. Registration of assessment protocols addresses two key factors important in the move toward OS-aware RRA. First is transparency, and second is mutual learning. Protocols include a description of the assessment, contextual factors, data sources, and indicators (and how they were calculated). Not included are individual identities and the specific evidence used. The collection of registered protocols is a searchable resource for others looking for inspiration in designing assessment approaches for OS contributions and/or RRA more broadly.

4.2. Interoperability & access layer

The interoperability across the diverse GraspOS Data Asset Sources is intended to be achieved following the recommendations outlined by relevant initiatives, with the RDA Working Group on Scientific Knowledge Graphs - Interoperability Framework (SKG-IF)[23] being the most important one. The SKG-IF aims to target the high fragmentation, heterogeneity and replication of scholarly information across different Scientific Knowledge Graphs (SKG) and reduce duplication of effort and capitalise on synergies and complementarity.

By comparing and putting to a common factor the modelled information contained in different SKGs publicly available such as Crossref, OpenAIRE, OpenCitations, and OpenAlex, the SKG-IF

- captures a set of core entities participating in scholarly communication processes such as research products, persons, grants, and organisations and their most relevant properties and relationships, and
- provides guidelines towards a pragmatic exchange of core information in a uniform way.

In its present version, the SKG-IF does not offer details about the exchange of information that is conducive for GraspOS tasks (e.g., indicators associated with research products), and therefore, it is our intention to extend it in order to accommodate such aspects that are not yet captured. The proposed extensions, of course, will be suggested back to the WG for consideration and possible inclusion in a newer iteration of the core model. However, it should be noted that, since the focus of SKG-IF is not on research assessment, it is highly likely that many of these extensions will not be incorporated into the basic model.

[23] SKG Interoperability Framework, https://skg-if.readthedocs.io/en/latest/index.html

At the moment, the modelled entities by SKG-IF are as follows:

- **Research products**, which may be of four types.
  - **Literature:** Intended for reading by humans (article, thesis, peer-review, blog posts, books, reports, patents, etc.)
  - **Research data:** Self-contained, persistently identified digital assets intended for processing (e.g., files containing: tables, metadata collections, dumps; persistent dynamic queries to scientific databases)
  - **Research software:** (definition from RDA WG) Research Software includes source code files, algorithms, scripts, computational workflows and executables that were created during the research process or for a research purpose. Note that software components (e.g., operating systems, libraries, dependencies, packages, scripts, etc.) that are used for research but were not created during or with a clear research intent should be considered software in research and not Research Software.
  - **Other products:** any digital asset, uniquely identified, whose nature does not fall into the first three types.
- **Persons**, an entity representing an individual who is involved in the creation, publication, and dissemination of Research products. A Person can be an author, a reviewer, an editor, a publisher, a researcher, or any other stakeholder involved in the scholarly communication process.
- **Organisations**, an entity that represents academic institutions, research centres, funders, or any other institutions taking part in the research process.
- **Venues**, an entity that models a publishing "gateway" used by a Person to make their Research products available to others.
- **Data sources**, an entity representing a service where published material (metadata and files) are stored, preserved, and made discoverable and accessible. A data source is described by the EOSC Profile for data sources.
- **Grants**, an entity that describes funding awarded to a Person or an Organisation by a funding body. These bodies, both public and private, can be funders, foundations, governments, agencies or institutions.
- **Topics**, an entity describing the scientific disciplines, the subjects and the keywords potentially relevant to a Research product.

In the context of GraspOS, the providers of the federated sources will build upon these data models and specifications to create a set of guidelines for onboarding data asset sources to the GraspOS infrastructure (and the respective dataspace in the future) and will work on making the required changes from their side and deploying the required API endpoints so that their sources will be federated.

4.3. Enrichment services

BIP! Citation Classifier

The BIP! Citation Classifier aims at extending the citation data aggregated by BIP! from different sources (e.g., Crossref, OpenCitations, OpenAIRE Graph) to calculate the indicators included BIP! DB; to this end, it first focuses on identifying citations that are not reported in those sources. In particular, since conferences and workshops are very important for the Computer Science domain and since it is common for Computer Science conferences and workshops to not assign DOIs to their articles, it collects OS publications lacking DOIs from DBLP, the most widely known bibliographic database for publications from Computer Science, and after performing text analysis techniques extracts citation information directly from the respective manuscripts. The output data are provided as an open resource on Zenodo (https://zenodo.org/record/8163673). The tool also supports a naive way of automatically annotating citations based on their intent, a functionality that is expected to be improved and extended in the context of the GraspOS project.

BIP! Services - Ranker

BIP! Services is a set of services that form a platform offering scientific literature exploration and research assessment services leveraging advanced citation-based impact indicators on top of scholarly knowledge graphs. BIP! aggregates citation data from Crossref, OpenCitations, and the OpenAIRE Graph, constructing a citation network that contains more than 130 million research products (articles and datasets). A set of citation-based indicators are, then, computed on top of this network in a scalable manner, capturing distinctly different aspects of scientific impact, such as popularity (current impact), influence (overall impact), and impulse (initial momentum). The aforementioned impact indicators are calculated by the BIP! Ranker component[24] and, after their calculation, they are used to (a) rank search results in BIP! Finder (https://bip.imsi.athenarc.gr/search), i.e., offering impact-based ranking

[24] BIP!-Ranker, https://github.com/athenarc/Bip-Ranker

functionalities to support customised literature exploration scenarios, and (b) calculate researcher-level indicators to be included in researcher profile pages, in BIP! Scholar (https://bip.imsi.athenarc.gr/scholar).

Within the context of GraspOS, it is planned to offer aggregate metrics to research groups and institutions and to provide additional metrics that will cover usage data and the level, impact, collaboration, and timeliness in OS practice for researchers, groups, and institutions.

EC KIP OS indicators

A suite of advanced software components for indicators on impact and collaboration of OA-based research outputs on citation and network analysis (with emphasis where possible on timeliness): Field-Weighted Citation Impact - FWCI scores, Collaborative Index (CI), Degree of collaboration (DC), Collaborative coefficient (CC).

OpenAIRE Broker

The OpenAIRE Broker is a subscription-notification service that enables content providers (repositories, CRIS systems, aggregators, knowledge graphs, publishers) to enrich their content with additional metadata. It utilises the OpenAIRE Graph and has a key role in the decentralisation and local re-use of all metadata available.

Within the context of GraspOS, we plan to extend the current data model to cover additional research results, relationships between products, classifications, and other metrics included in metadata records.

OpenAIRE IIS text-mining modules

An information inference service that enriches scholarly data with automatically inferred metadata, based on a flexible big data processing pipeline supporting full-text and metadata mining. Current modules include citation extraction (article-data, article-software, product-grant, product-organisations); subject inference (Frascati and SDG); community context.

Within the context of GraspOS, we plan to extend it with methods for the completeness of metadata (e.g., resource types, subjects, licensing typology) and, in particular for the pilots, deeper classifications (> Frascati Level 4) to be used in capturing discipline-specific characteristics.

OpenAIRE metadata validator

A rule-based service that provides metrics on the compatibility with the OpenAIRE Guidelines and on the FAIRness of the data sources (repositories, journals, CRISs) based on the metadata of the research output.

Within the context of GraspOS, we plan to extend it with configurations for custom and domain-specific FAIRness metrics at the level of the individual records and average-based metrics for data sources, funders, and institutions. Publish as a stand-alone service (now embedded in PROVIDE Dashboard).

SCRE pipeline

A general and configurable AI-assisted content acquisition and processing pipeline for documents and projects that aggregates and performs data cleaning and semantic processing.

Within the context of GraspOS, we plan to extend the pipeline with methods for the completeness of metadata for SSH. Upgrade so outputs match the specifications from RDA IG Scientific Knowledge Graphs.

Semantic citation classifier

A software that performs the automatic annotation of in-text citations in academic papers provided in PDF. It works by applying two steps, described as follows.

The first step is *PDF Parsing*. The software analyses the PDF paper provided as input and extracts its basic bibliographic metadata (mainly the authors and the title), all the bibliographic references with all its metadata (authors, year of publication, title, venue, identifiers) marked up, the citation sentences that contain in-text reference pointers (i.e., the textual device used to refer to bibliographic references such as "[3]" and "(Doe et al., 2023)"), and other structural information such as sections, when possible. All these data will be returned as an RDF dataset compliant with the OpenCitations Data Model (OCDM, https://opencitations.net/model) (Daquino *et al.*, 2020).

The second step is the *Citation Function Classification*. Using an approach that mixes up different technologies, such as Large Language Models (LLM) and Knowledge Graph Embeddings (KGE), the software uses the RDF dataset describing citation information of the input PDF to classify the semantics emerging from each citation sentence that will be used for characterising the function of the citation defined by the authors of the citing paper (i.e., the input PDF) by means of the related in-text reference pointer. The citation functions returned by the software will be a subset of those defined in the Citation Typing Ontology (CiTO, http://purl.org/spar/cito) (Peroni & Shotton, 2012). The output of this step will be the same RDF dataset provided as input, enriched with the specification of the citation functions associated with each citation.

4.4. Monitoring services

BIP! Scholar

BIP! Scholar is a platform that enables researchers to create researcher profile pages representing their research activities and highlighting different aspects of their research career. BIP! Scholar leverages data from scholarly knowledge graphs (OpenAIRE Graph, Crossref, OpenCitations) and ORCID to display custom reports containing a variety of researcher-level RRA indicators for researchers capturing different aspects of their performance (e.g., productivity, impact, career stage) while considering the different roles of researchers in the respective works (according to the CRediT taxonomy). The calculated impact indicators are computed on a citation network that consists of more than 130 million research products (articles and datasets) and more than 1.5 billion citations among them; access to those records is also available via an Open Access API (https://bip.imsi.athenarc.gr/site/data).

Within the context of GraspOS, detailed reports on researchers' OS practice (level and impact) will be produced. BIP! Scholar will offer such reports, supported by appropriate (intuitive) visualisations, both on the researcher level as well as on groups of researchers. Finally, additional concepts that can facilitate RRA (e.g., narrative CVs) will be integrated into BIP! Scholar.

EOSC accounting for services

A tool that collects service metrics such as Virtual Access Metrics and aggregates them according to EOSC/community rules. It would be used as a data source for impact metrics for the services as part of the federated metrics infrastructure.

Within the context of GraspOS, it is planned to extend with an add-on and APIs to provide aggregate-level metrics and reports.

OpenAIRE institutional monitor dashboard

A service[25] built on the OpenAIRE Graph providing monitoring services for the research output of stakeholders (i.e., RPOs, RFOs, researchers). The service offers OS statistics focusing on institutions and their subunits, funders, and research initiatives. It is highly configurable to accommodate mix & match indicators and metrics.

Within the context of GraspOS, it is planned to be extended with additional profiles (researchers) and indicators. The existing administration backend will be enhanced with pre-set templates, and support others by building additional ones tailored to their needs.

OpenAIRE Researcher Profile

OpenAIRE Researcher Profile is a new service (to be developed in the context of the GraspOS projects) that will extend the functionality of OpenAIRE EXPLORE[26], the discovery portal of scholarly works that is built upon the OpenAIRE Graph. OpenAIRE Researcher Profile will provide profile pages for researchers, where their research products, such as publications, data, and software will be displayed. The profile pages will also include statistics and indicators that summarise the researchers' aggregated impact, productivity, and compliance to OS practices. OpenAIRE Researcher Profile will enable researchers to increase their visibility, track their performance, and demonstrate their adherence to OS principles.

Open Science Observatory

The Open Science Observatory[27] (OSO), is an online platform that offers rich visualisations of various OS aspects in Europe. It draws data from the OpenAIRE Graph, a comprehensive and open scholarly communication graph, and other public data sources. The OSO allows end

[25] OpenAIRE MONITOR, https://monitor.openaire.eu

[26] OpenAIRE EXPLORE, https://explore.openaire.eu

[27] Open Science Observatory (OSO), https://osobservatory.openaire.eu

users to explore and compare the impact, productivity, and compliance of Horizon 2020 and Horizon Europe, and other funded research projects to OS practices. It follows a top-down methodology for deriving indicators based on high-level monitoring targets and employs metrics that can measure the openness of research output (publications, data, software or other research products) on various aspects (e.g., gold/green/fair) and the regional or thematic distributions (at EU, Country and Repository-level). It aims to provide services to funding agencies, policymakers, research organisations and researchers and help them assess different dimensions of OS research. The OSO offers interoperable data to the EOSC Observatory, offering analytics from data sources in Europe that may serve for evidence-based policy.

EOSC Observatory

The EOSC Observatory is a policy intelligence tool with the aim of monitoring policies, practices, and impacts related to OS and the European Open Science Cloud (EOSC). The EOSC observatory essentially tracks the EOSC readiness of Member States and Associated Countries (MS/AC), including their contributions, investments, and implementation of EOSC. The observatory consists of a back-end for running surveys and analysing responses, as well as a front-end for visualising and exploiting collected data. Another section of the EOSC Observatory is completed by the National Open Access Desks (NOADs) in the upcoming "EOSC Observatory Country Pages", which provide the updated state-of-play, key statistics, and relevance contacts and links for EOSC in each country.

4.5. Data registry

This component, which will be developed in the context of the GraspOS project, will allow users to be informed about and understand the contents within the infrastructure (and the dataspace to be subsequently developed based on this). The data registry will be developed by adapting and/or extending an open-source software, and it will collect metadata and access information for the onboarded data sources. More specifically, apart from a set of basic metadata related to each data source (e.g., its title, owners, licence), the GraspOS data registry will also contain information about the location/address of the respective GraspOS API deployment, enabling the programmatic access to the data source contents.

5. Conclusions

This deliverable introduces the first version of the GraspOS architecture.

The GraspOS Federated Infrastructure is designed to support several functional goals, including the implementation of OS-aware RRA processes, the promotion of OS and collaboration, and the development of new research assessment indicators and methods. The infrastructure is also designed to be flexible and adaptable, allowing for the integration of new data sources and the development of new use cases and services.

The GraspOS infrastructure is composed of a set of federated Data Asset Sources providing access to and delivering research products metadata, pre-calculated research performance indicators, as well as unstructured data such as scientific texts, narrative CVs and data on OS practices.

These sources are flanked with a set of federated monitoring and enrichment services, which can be used to handle the transformation, manipulation, and analysis of data within the infrastructure, and present and visualise research assessment evidence and indicators at different levels of granularity (e.g., researcher, organisation, country) to support OS-aware RRA and to monitor the uptake and impact of OS. A Data Registry documenting the federated resources complements the infrastructure.

The communication across all the components federated in the GraspOS infrastructure is granted by an interoperability-by-design approach adhering to a set of guidelines drawn by research-community-driven initiatives.

6. References

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. Science, 159(3810), 56-63. https://doi.org/10.1126/science.159.3810.56

Strinzel, M., Brown, J., Kaltenbrunner, W., de Rijcke, S., & Hill, M. (2021). Ten ways to improve academic CVs for fairer research assessment. Humanities and Social Sciences Communications, 8, 251. https://doi.org/10.1057/s41599-021-00929-0

Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. (2020) The Hong Kong Principles for assessing researchers: Fostering research integrity. PLoS Biol 18(7): e3000737. https://doi.org/10.1371/journal.pbio.3000737

International Network of Research Management Societies - Research Evaluation Group (2023). The SCOPE Framework. https://doi.org/10.26188/21919527.v1

Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., ... & Zumstein, P. (2020). The OpenCitations data model. In International Semantic Web Conference (pp. 447-463). https://doi.org/10.1007/978-3-030-62466-8_28

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Journal of Web Semantics, 17, 33-43. https://doi.org/10.1016/j.websem.2012.08.001