publications

Case Report

FAIR as a Journey: Lessons Learned from Building the GoTriple Discovery Platform for Social Sciences and Humanities

Luca De Santis

Net7 Srl, 56124 Pisa, Italy; desantis@netseven.it

Abstract: This report describes the experience in implementing the FAIR principles for the GoTriple Discovery Platform for Social Sciences and Humanities (SSH). It shows how adherence to FAIR should be considered as a continuous process throughout the entire lifespan of any information management system, including GoTriple, rather than a static goal with decisions only made at the design time. This report presents an introduction highlighting the importance of the FAIR principles, indicating how they can be assessed in data management systems. Then, the GoTriple case is presented, with a general overview of this discovery platform before describing some of the implemented practices in support of FAIR. The Discussion Section shows, on the one hand, some virtuous reuse of GoTriple data, together with one major pitfall in the platform's FAIR implementation. In this sense, this report serves as a case study that can offer insights and actionable advice for those implementing information systems aligned, from the very outset, with the FAIR principles.

Keywords: FAIR principles; open science; Social Sciences and Humanities (SSH); information and data management systems; discovery platforms for SSH research; OPERAS services

check for updates

Citation: De Santis, L. FAIR as a Journey: Lessons Learned from Building the GoTriple Discovery Platform for Social Sciences and Humanities. *Publications* **2024**, *12*, 26. https://doi.org/10.3390/publications12030026

Academic Editors: Iva Grabarić Andonovski, Nikolina Peša Pavlović, Jadranka Stojanovski and Andrew Kirby

Received: 3 May 2024
Revised: 9 August 2024
Accepted: 28 August 2024
Published: 29 August 2024

1. Introduction: FAIR as a Journey

To obtain the maximum value from information, it is essential that both the data and associated metadata respect the characteristics of being Findable, Accessible, Interoperable, and Reusable (henceforth, FAIR). FAIR was introduced for the first time in 2016 with the seminal article by Wilkinson and colleagues [1]. Their *FAIR Guiding Principles for scientific data management and stewardship* highlighted the "urgent need to improve the infrastructure supporting the reuse of scholarly data".

FAIR is now a pillar of open science, albeit the road to reach FAIRness is paved with obstacles and difficulties.

Moreover, while it is advisable that information management systems are designed from the very beginning to be FAIR, it is common that adjustments are performed along the way to improve or even implement technical solutions to respect or increase their FAIRness compliance, as long as new data are ingested and managed.

Therefore, it makes perfect sense to think of "FAIR as a journey", as specified in the *Recommendations on FAIR metrics for EOSC* of the European Commission [2], and not simply as a static requirement. It should be seen as a road to take and never steer away from. In [2], it is claimed that "...if FAIR is not seen as a continuum, we risk alienating communities who are not well advanced in sharing their data in a FAIR way, or indeed in data sharing at all. We also risk losing advanced communities for whom the effort to attain optional indicators outweighs the expected benefit".

Therefore, it is advisable to approach FAIR using a "continuous improvement" process, like those used by modern agile project management systems such as SCRUM [3] for software development or FitSM [4] for IT service management.

Assessing FAIR compliance has become possible thanks to the guidelines specified in the previously mentioned *The FAIR Guiding Principles* article [1] and in *Recommendations on FAIR metrics for EOSC* of the European Commission [2]. The *FAIR Data Maturity Model*

Specification and Guidelines of the FAIR Data Maturity Model Working Group [5] also provide actionable indicators to measure FAIRness.

The former presents the 15 principles that must guide every implementation of FAIR.

The European Commission's document includes seven recommendations on the definition and implementation of metrics for FAIR data. While focusing on the context of the European Open Science Cloud (EOSC), that report provides useful indications for anyone interested in the FAIRness of data, including the already cited concept of considering "FAIR as a journey" (Recommendation 2.2). Finally, the latter provides a complete list of indicators for verifying adherence to the FAIR principles. It consists of 41 criteria that easily allow us to assess FAIR compliance both for data and for their associated metadata, specifying for each one of them three possible levels of implementation priority (essential, important, useful).

By analysing these principles, it is evident how technical support from IT specialists (in-house developers/sysadmins or external consultants/system integrators) is mandatory to ensure the success of the FAIRness strategy. Think, for example, of the indicators RDA-F4-01M ("Metadata is offered in such a way that it can be harvested and indexed") or RDA-A1-04D ("Data is accessible through standardised protocol") in [5]. It is evident that they cannot be fully achieved without strict collaboration with the technical team in charge of the development, operation, and maintenance of an information management system. At the same time, FAIR is a real philosophy of data management that cannot be just dismissed as a set of technical issues to implement or solve.

The experience of developing GoTriple, a discovery platform for the Social Sciences and Humanities, serves as a good example of how FAIR must be intended as a continuous journey, of which this report represents a sort of logbook. Here, the solutions provided, the good practices put in place, and also the pitfalls of the implementation of FAIR for GoTriple are presented. Many of these strategies proved to be quite solid and stood real tests, with virtuous examples of reuse that are described herein. At the same time, other choices turned out to be weak and quite expensive to correct.

This report is intended as a guide for those planning to develop a data management system, offering hints that, if followed, can make the road to FAIR less arduous. In the following section (Section 2), the experience with FAIR for the GoTriple.eu website is described. After a general presentation of GoTriple (Section 2.1), the actual solutions implemented to adhere to the FAIR principles are described (Sections 2.2–2.5), with explicit references to the "FAIR Guiding Principles", as defined in Wilkinson et al.'s article [1], to facilitate understanding. Section 3 shows how the FAIR principles have been translated into practice, highlighting both the effective reuse of GoTriple data and a major pitfall in the platform's FAIR implementation. The conclusions provide a summary and suggest possible directions for GoTriple's implementation roadmap, particularly concerning FAIR compliance.

2. Detailed Case Description: FAIR in GoTriple

2.1. Introducing GoTriple

GoTriple (https://gotriple.eu, accessed on 20 August 2024) is a multilingual discovery platform for the Social Sciences and Humanities. It is the main outcome of the TRIPLE research project [6], which received funding from the European Union's Horizon 2020 Research and Innovation action (funding scheme INFRAEOSC-02-2019 "Prototyping new innovative services"). The project, whose full name is "Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration", ran between October 2019 and March 2023. It was coordinated by CNRS and featured 22 partners from 15 different European countries.

GoTriple provides a central access point that allows users to explore, find, access, and reuse materials such as articles, datasets, project descriptions, and author profiles at the European scale.

It is one of the Discovery Services of OPERAS [7], the Research Infrastructure that supports open scholarly communication in the Social Sciences and Humanities in the European Research Area.

At its heart, GoTriple has a search engine whose indexes are fed by a configurable harvesting and processing pipeline, which continuously imports and processes publication and project metadata from multiple sources (about 1400), including large aggregators (BASE, DOAJ, OpenAire, Isidore) and national repositories alike (Hrčak, Biblioteka Nauki, ZRC Sazu, EKT, Recyt or Pombaline).

Recently, the support of MARC21 XML data sources has been implemented in GoTriple, which allowed for the ingestion of metadata from over 1.7 million documents in the German National Library (DNB) [8].

Innovative services are integrated into the platform to improve user experience with personalised recommendations, interactive visualisations, and the possibility to use a web annotation tool to take notes on material found in GoTriple.

Finally, users can register to the platform to create a personal profile, claim ownership of the documents published in the indexes, and find and connect with other SSH authors and researchers.

2.2. Findability

The requirement of making a digital asset "findable" is expressed by the four principles mentioned in Table 1.

Table 1. FAIR guiding principle: findability.

To Be Findable	
F1.	(meta)data are assigned a globally unique and persistent identifier
F2.	data are described with rich metadata
F3.	metadata clearly and explicitly include the identifier of the data they describe
F4.	(meta)data are registered or indexed in a searchable resource

The first one (F1) requires that each data resource is assigned a unique and persistent identifier. GoTriple respects this requirement by identifying each resource with a Uniform Resource Identifier (URI), which is the URL of its landing page. From a theoretical viewpoint, this requirement is fully respected, as this identifier is unique and persistent (at least as long as GoTriple exists). As a matter of fact, this design choice proved to be a pitfall in the global GoTriple implementation, which will be commented on in depth in Section 3 (Discussion: FAIR Principles in Practice within GoTriple).

The F2 principle indicates the importance of describing data assets with rich metadata. In this sense, GoTriple fully respects this requirement by providing a detailed vision for its data model, whose main structures are shown in Figure 1. It consists of three main asset types, i.e., document, profile, and project, which were initially designed by using standard ontologies like Schema.org and SIOC to describe their attributes.

It reflects the type of metadata that can be harvested from the available data sources. In particular, every publication creates an entry in the document index, while its authors are stored in the profile index, with a link between them. Projects, on the other hand, have so far been harvested from sources without meaningful associated research publications, or where it was difficult to determine whether the mentioned persons are simply administrators or actual SSH researchers. This explains the lack of relationships between the project index and the document and profile indexes.

The F3 principle requires that "metadata clearly and explicitly include the identifier of the data it describes". All original identifiers are retrieved during the harvesting process of GoTriple content and shown in their descriptive pages of the website, as shown in Figure 2.

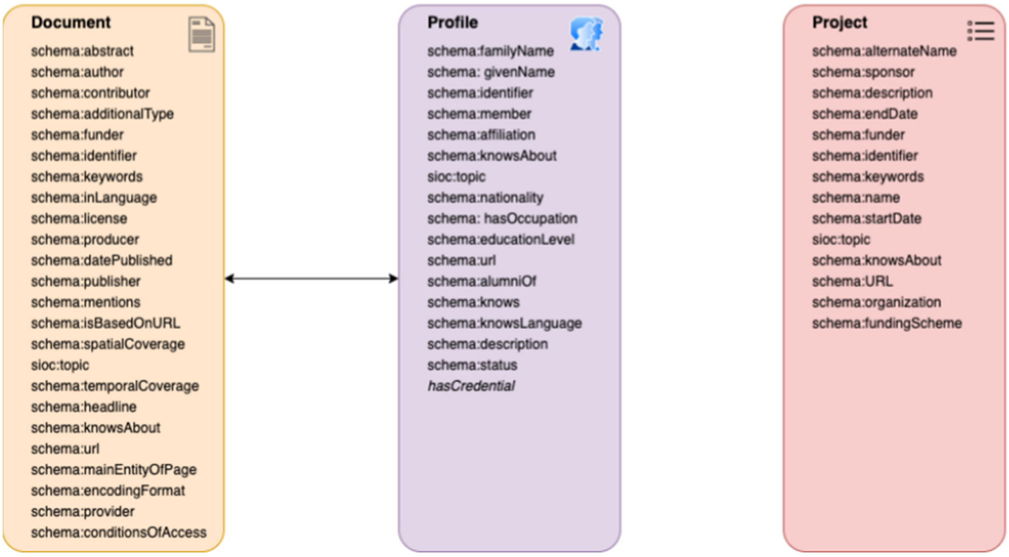
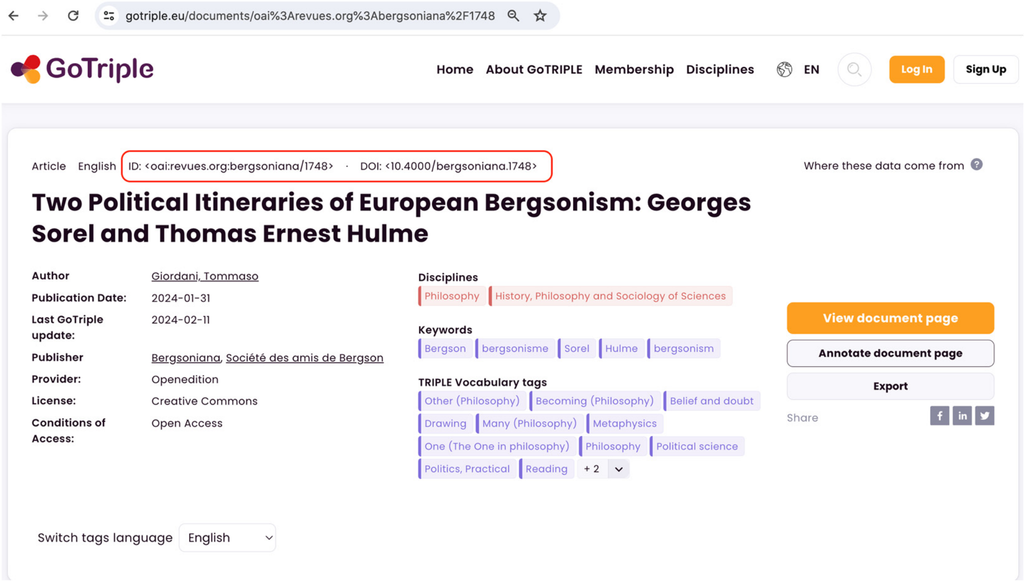


Figure 1. TRIPLE data model.



Article English ID: <oai:revues.org:bergsoniana/1748> · DOI: <10.4000/bergsoniana.1748>

Where these data come from

Two Political Itineraries of European Bergsonism: Georges Sorel and Thomas Ernest Hulme

Author	Giordani, Tommaso
Publication Date:	2024-01-31
Last GoTriple update:	2024-02-11
Publisher	Bergsoniana, Société des amis de Bergson
Provider:	Openedition
License:	Creative Commons
Conditions of Access:	Open Access

Disciplines
Philosophy | History, Philosophy and Sociology of Sciences

Keywords
Bergson | bergsonisme | Sorel | Hulme | bergsonism

TRIPLE Vocabulary tags
Other (Philosophy) | Becoming (Philosophy) | Belief and doubt | Drawing | Many (Philosophy) | Metaphysics | One (The One in philosophy) | Philosophy | Political science | Politics, Practical | Reading | + 2

Switch tags language English

Figure 2. GoTriple document page. The original identifiers of the resource are maintained and shown.

Finally, F4 requires that a searchable index is required to access the resources easily. As indicated, search is one of the main features of GoTriple, which offers an intuitive faceted interface for further filtering the list of results.

2.3. Accessibility

Accessibility demands that data and metadata are retrievable by both humans and machines through well-defined, open, and universally implementable protocols. Table 2 shows in detail the accessibility requirements.

All GoTriple data are accessible via the open and standard HTTP protocol (principles A1. and A1.1), providing interfaces both for humans (the website GoTriple.eu) and software programs, the latter in the form of REST APIs, accessible publicly or only after previous authentication (A1.2). APIs' documentation is available through a link on the GoTriple website [9]. GoTriple's document metadata are also accessible for harvesting through the standard OAI-PMH [10] protocol. Finally, in its local indexes, GoTriple copies the metadata of the original assets, guaranteeing their preservation even if the data in the original source, for example, the full text of an article, are no longer accessible (A2.).

Table 2. FAIR guiding principle: accessibility.

To Be Accessible	
A1.	(meta)data are retrievable by their identifiers using a standardised communications protocol
A1.1	the protocol is open, free, and universally implementable
A1.2	the protocol allows for an authentication and authorisation procedure, where necessary
A2.	metadata are accessible, even when the data are no longer available

Beyond the FAIR guiding principles of accessibility indicated above, it is important to highlight a strategic decision taken in GoTriple to enhance users' access to data. As a multilingual platform, with articles written in many different languages, to improve readability, GoTriple always provides an English translation of the text of its assets, by resorting, when necessary, to eTranslation [11], an automatic neural machine translation service provided by the European Commission.

2.4. Interoperability

Interoperability in FAIR is based on the assumption that to facilitate the exchange of information amongst systems, data and metadata should be defined by following standard formats, using shared vocabularies and providing references to other resources. Its guiding principles are presented in Table 3.

Table 3. FAIR Guiding Principle: interoperability.

To Be Interoperable	
I1.	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2.	(meta)data use vocabularies that follow FAIR principles
I3.	(meta)data include qualified references to other (meta)data

Over time, the GoTriple data model has been formally defined through the TRIPLE Ontology [12] (see principle I1.). This allowed for not only providing a semantic web-compliant, machine-understandable description of the three main asset types of the platform but also the controlled vocabularies used to specify the values of the "license", "conditions of access", "document type", and "discipline" attributes (see principles I2.).

The former two simply provide the list of the admitted identifiers used in GoTriple.

"Document type" on the other hand provides a linked data description of the 20 admitted types by linking them to the corresponding resource types of the COAR Controlled Vocabularies for Repositories [13].

The "disciplines" controlled vocabulary is again a linked data resource that uses the SKOS ontology [14] to provide a formalisation of the SSH domain in 27 fields of study. This formalisation is one of the outcomes of the TRIPLE project and exploits the vision of a former EU-funded research project (MORESS [15]). In this vocabulary, each discipline is linked to one or more corresponding concepts of other widely known classification systems, including the Library of Congress Subject Headings, Wikidata, and the Dewey Decimal Classification. Finally, the vocabulary is also multilingual, with concepts available in English, Italian, French, and German.

GoTriple also makes use of automatic classification systems by which all the content ingested is assigned to one or more disciplines and "tagged" with concepts belonging to TRIPLE vocabulary [16]. The latter consists of a rich controlled vocabulary of over 3370 linked data SSH-related concepts. This is another significant outcome of the TRIPLE project: it has been formalised in SKOS and provides translations in 11 languages.

The presence of a formal description of its assets, along with several linked data attributes, improves interoperability and enables better reuse of GoTriple's content (see principle I3.)

2.5. Reusability

Reusability, whose guiding principles are indicated in Table 4, ensures that data and associated metadata are well-documented and licensed in a clear and accessible manner, to enable their future use by others with minimal restrictions.

Table 4. FAIR guiding principle: reusability.

To Be Reusable	
R1.	meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1	(meta)data are released with a clear and accessible data usage license
R1.2	(meta)data are associated with detailed provenance
R1.3.	(meta)data meet domain-relevant community standards

GoTriple's entity descriptions proved to be complete and accurate (see R1.). As a result, the original metadata fetched from external sources normally can only match a part of the platform's data model. This is especially true for the majority of the harvested sources, where metadata are imported using the OAI-PMH harvesting protocol and described in Dublin Core, which provides only a limited set of metadata elements.

Licenses and conditions of access for using the imported metadata are preserved and republished on GoTriple using a controlled vocabulary (see Section 2.4), provided they were specified at the source (see R1.1).

A detailed explanation of the origin of the documents in the platform's index is also provided (R1.2). By clicking on a link on the presentation page of a document, the user can distinguish the attributes coming from the original source from those produced by the GoTriple processing pipeline through the curation and enriching phase.

Finally concerning R1.3, as mentioned previously, GoTriple's controlled vocabularies are carefully defined by linking them to widely recognised standards. It is also worth mentioning the possibility for users to export document metadata easily in a variety of standard formats, including BibTeX, JSON, and JSON-LD.

3. Discussion: FAIR Principles in Practice within GoTriple

In the previous section, GoTriple's compliance with the FAIR principles was described. Beyond the mere declaration of commitment, it is useful to see how all the choices made and presented above translate into practice. Most of these measures proved to be useful and beneficial for the research community at large, as they enabled data interoperability and reuse in several contexts.

The first example is provided by VERA [17], OPERAS' service for participatory research. VERA is a web-based platform that enables laypersons and researchers to collaborate on SSH-related citizen science initiatives. Projects in VERA are defined by a set of descriptive attributes, including "Academic subjects", for which GoTriple's disciplines have been used. Additionally, a VERA project can be automatically published on GoTriple (see the "EcoVoce" VERA project as an example [18]), and the use of this shared classification not only ensures its correct presentation on the Discovery Platform but also helps create connections with other initiatives in the same field of study.

GoTriple open APIs also facilitated the reuse of GoTriple's content on various occasions. For example, a team of the AGH University of Krakow, led by professor Mikołaj Leszczuk, used GoTriple APIs for research on identifying similar illustrations in scientific publications [19]. Using APIs, they collected over 5600 scientific papers for which their full texts in PDF were available. This enabled the team to extract almost 65,000 images, which helped them to develop and fine-tune the software methodology at the core of their research.

Finally, through a collaboration between the University of Pisa and the Italian company Net7 [20], an Artificial Intelligence-driven chatbot was implemented on GoTriple's content. In a similar way, the full texts of over 1000 articles of a specific SSH discipline were retrieved and used to implement an interactive assistant, able to respond to actual research questions and to provide references of the sources used.

It is also important to highlight that despite full compliance with FAIR principles, not all of GoTriple's design and implementation choices have proved to be valid. One pitfall worth mentioning is the implementation of Persistent Identifiers (PIDs). The importance of PIDs is emphasised by the fact that all related indicators in [5] are marked as "essential".

As mentioned, in GoTriple, a PID consists of the URL of the presentation page of an entity. This URL is created by merging GoTriple's domain name with the original identifier of the content from the remote source, which is usually. the first one returned during the harvesting phase. An example is provided below.

https://www.gotriple.eu/documents/<PRIMARY_ID>

For example, the PID of a GoTriple document is as follows:

https://www.gotriple.eu/documents/oai:revues.org:bergsoniana/1748, accessed on 20 August 2024.

In general, using URLs for IDs is a bad idea. Some research (see [21], for example) mentions that over a long period, a significant percentage of URLs (over 50%) become inaccessible. Also, linking an external identifier so strictly to GoTriple's ID is risky. In fact, when harvesting sources, it is common to receive multiple IDs for the same resource, and the first one taken is not always the one intended to be persistent. It might be an internal ID that can change over time, provided that, if present, the persistent ID (e.g., a DOI) is maintained.

Unfortunately, this consideration was realised too late in the implementation of GoTriple, after millions of records had already been ingested and published. Fixing this problem by creating PIDs with a more solid logic for both new and existing content is one of the goals for the next phase of GoTriple's implementation. One interesting approach to consider is the one used by OpenAIRE for the entities of its graph [22]. Here, the PIDs are created by merging an identifier of the source and a hash code generated by processing the content's original identifier, privileging, for this operation, those known to be persistent (DOI, Handle System identifiers, ISBN). Of course, this operation in GoTriple must be planned with great care in order to maintain backward compatibility with the current platform's identifiers.

Finally, it must be noted that most of the effort in designing and implementing GoTriple's FAIR support has been devoted to documents that represent the majority and possibly the most valuable content of the Discovery Platform. While the TRIPLE Ontology has recently been expanded to support projects and profiles, some other FAIR-related features, such as the accessibility of content through standard harvesting protocols like OAI-PMH, are currently only available for documents.

4. Conclusions

The journey towards FAIR must be intended as a continuous endeavour, and the GoTriple case study is exemplary in this regard. The current development plan for the platform aims not only to enhance existing achievements (such as fixing Persistent Identifiers and increasing FAIRness for projects and profiles) but also to improve the quality of metadata and the general alignment of GoTriple vocabularies with other relevant SSH-related ontologies.

This effort is being pursued under the ATRIUM research project [23], funded by the European Union under the call HORIZON-INFRA-2023-SERV-01. Here, some of the original partners of the TRIPLE project (Coimbra University, Foxcub, IBL-PAN, Net7) collaborate with the OPERAS infrastructure to advance GoTriple, also enhancing its FAIRness.

Funding: This research was funded by the European Union, grant agreement numbers 863420 (TRIPLE [6]) and 101132163 (ATRIUM [22]).

Data Availability Statement: Data sharing not applicable—no new data generated.

Acknowledgments: The author wants to thank the Pubmet2023 team at the University of Zadar, in particular, Jadranka Stojanovski and Drahomira Cupar, whose friendship honours him.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*. [CrossRef] [PubMed]
2. Genova, F.; Aronsen, J.M.; Beyan, O.; Harrower, N.; Holl, A.; Hooft, R.W.; Principe, P.; Slavec, A.; Jones, S. *Recommendations on FAIR Metrics for EOSC*; Publications Office of the European Union, 2021. Available online: https://op.europa.eu/en/publication-detail/-/publication/ced147c9-53c0-11eb-b59f-01aa75ed71a1/language-en (accessed on 1 April 2024). [CrossRef]
3. Scrum.org Web Site. Available online: https://www.scrum.org (accessed on 1 April 2024).
4. FitSM Web Site. Available online: https://www.fitsm.eu/ (accessed on 1 April 2024).
5. FAIR Data Maturity Model Working Group. *FAIR Data Maturity Model. Specification and Guidelines*; Zenodo.org, 2020. [CrossRef]
6. TRIPLE Project Web Site. Available online: https://project.gotriple.eu/ (accessed on 1 April 2024).
7. OPERAS Web Site. Available online: https://operas-eu.org/ (accessed on 1 April 2024).
8. Deutsche Nationalbibliothek (DNB) Web Site. Available online: https://www.dnb.de/ (accessed on 1 April 2024).
9. De Santis, L. *TRIPLE Deliverable: D6.6 API's Development-RP3*; Zenodo.org, 2022. [CrossRef]
10. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Available online: https://www.openarchives.org/pmh/ (accessed on 1 April 2024).
11. eTranslation. Available online: https://commission.europa.eu/resources-partners/etranslation_en (accessed on 1 April 2024).
12. TRIPLE Ontology. Available online: https://www.gotriple.eu/ontology/triple (accessed on 1 April 2024).
13. COAR Controlled Vocabularies for Repositories: Resource Types 3.1. Available online: https://vocabularies.coar-repositories.org/resource_types/ (accessed on 1 April 2024).
14. SKOS—Simple Knowledge Organization System. Available online: https://www.w3.org/2009/08/skos-reference/skos.html (accessed on 1 April 2024).
15. MORESS—Mapping of Research in European Social Sciences and Humanities. Available online: https://cordis.europa.eu/project/id/HPSE-CT-2002-60060 (accessed on 1 April 2024).
16. TRIPLE Vocabulary. Available online: https://www.semantics.gr/authorities/vocabularies/SSH-LCSH/vocabulary-entries?language=en (accessed on 1 April 2024).
17. VERA—Virtual Ecosystem for Research Activation. Available online: https://vera.operas-eu.org/ (accessed on 1 April 2024).
18. EcoVoce Project in GoTriple. Available online: https://gotriple.eu/projects/vera:82-ecovoce (accessed on 31 July 2024).
19. Leszczuk, M.; Dziula, P. Implementation of Software for Searching Similar Illustrations in Scientific Publications. In Proceedings of the Opening Collaboration for Community-Driven Scholarly Communication (OPERAS2024), Zadar, Croatia, 24 April 2024; Zenodo.org. 2024. [CrossRef]
20. Bertozzi, A.; Abbamonte, M.L.; Abaza, A.; De Santis, L. From Words to Search: GoTriple's AI ChatBot for Efficient Research Engagement. In Proceedings of the Opening Collaboration for Community-Driven Scholarly Communication (OPERAS2024), Zadar, Croatia, 24 April 2024; Zenodo.org. 2024. [CrossRef]
21. Stojanovski, J. PIDs in the SSH—Current state and upcoming challenges. In Proceedings of the TRIPLE Booksprint—The role of Open Metadata in the SSH Scholarly Communication, Konstancin-Jeziorna, Poland, 7–9 September 2022.
22. OpenAIRE Graph Documentation. PIDs and Identifiers. Available online: https://graph.openaire.eu/docs/data-model/pids-and-identifiers (accessed on 1 April 2024).
23. Advancing fronTier Research in the Arts and hUManities (ATRIUM) Web Site. Available online: https://atrium-research.eu/ (accessed on 1 April 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.