Case Report

FAIR as a Journey: Lessons Learned and Takeaways from Building the GoTriple Discovery Platform for SSH

Luca De Santis *

Posted Date: 13 May 2024

doi: 10.20944/preprints202405.0748.v1

Keywords: FAIR Principles; Open Science; Social Sciences and Humanities (SSH); Information and Data Management Systems; Discovery Platforms for SSH Research; OPERAS Services

Case Report

FAIR as a Journey: Lessons Learned and Takeaways from Building the GoTriple Discovery Platform for SSH

Luca De Santis

Net7 Srl; desantis@netseven.it

Abstract: This report describes the experience in implementing the FAIR principles for the GoTriple Discovery Platform for the 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 design time. The 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

1. Introduction: FAIR as a Journey

FAIR is a pillar of open science. As a concept it has been introduced for the first time in the seminal article of 2016 by Wilkinson et al. (*The FAIR Guiding Principles for scientific data management and stewardship*) [1], in which it was highlighted the "urgent need to improve the infrastructure supporting the reuse of scholarly data".

To get the maximum value from digital information is in fact essential that data and their associated metadata respect the characteristics of being Findable, Accessible, Interoperable and Reusable. While this might sound obvious, 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 done along the way, to improve or even implement technical solutions to respect or increase their FAIRness compliance, as long as new data is 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. It must be approached by using a "continuous improvement" process, typical of modern agile project management systems, like SCRUM [3] for software development or FitSM [4] for service management.

Assessing and measuring the compliancy to FAIR has become possible, thanks to the guidelines specified in the already mentioned *The FAIR Guiding Principles* article [1], in *Recommendations on FAIR metrics for EOSC* of the European Commission [2], and to the actionable indicators specified in the *FAIR Data Maturity Model. Specification and Guidelines* of the FAIR Data Maturity Model Working Group [5].

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

 doi:10.20944/preprints202405.0748.v1

The European Commission's document includes 7 recommendations on the definition and implementation of metrics for FAIR data: while focusing on the context of the European Open Science Cloud (EOSC), this report provides useful indications for anyone interesting 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 the adherence to the FAIR principles. It consists of 41 criteria that easily allow 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 to 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 the 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, represents a good example of how FAIR must be really intended as a continuous journey, of which this report represents a sort of logbook. Here, the solutions provided, the good practices put in place but 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.

In this sense this report is also intended to be a guide for those who are planning to develop a data management system, with hints that, if followed, would make the road to FAIR a less arduous journey.

2. Detailed Case Description: FAIR in GoTriple

In this section the experience done in respect with FAIR for the GoTriple.eu web site is described. After a general presentation of GoTriple (2.1), the actual solutions put in place to respect the FAIR principles are described (2.2, 2.3, 2.4 and 2.5): the "FAIR Guiding Principles", as defined in the Wilkinson et al.'s article [1] are explicitly mentioned to facilitate the presentation of the work done for GoTriple.

2.1. Introducing GoTriple

GoTriple 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 authors profiles at 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, that continuously imports and processes publications and projects metadata from multiple sources (about 1.400), including large aggregators (BASE, DOAJ, OpenAire, Isidore) and national repositories alike (Hrčak, Biblioteka Nauki, ZRC Sazu, EKT, Recyt or Pombaline).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2024 doi:10.20944/preprints202405.0748.v1

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

Innovative services are integrated into the platform to improve the user experience with personalized 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, to claim the ownership of the documents published in the indexes, and to 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. The FAIR Guiding Principles: 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 it describes
F4.	(meta)data are registered or indexed in a searchable resource

The first one (F1) requires that each data resource has 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 has been 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 that will be commented in depth in the 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, Document, Profile and Project, which were initially designed by using standard ontologies like Schema.org and SIOC for describing their attributes.

 doi:10.20944/preprints202405.0748.v1

Figure 1. TRIPLE Data Model.

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 web site, as shown in Figure 2.

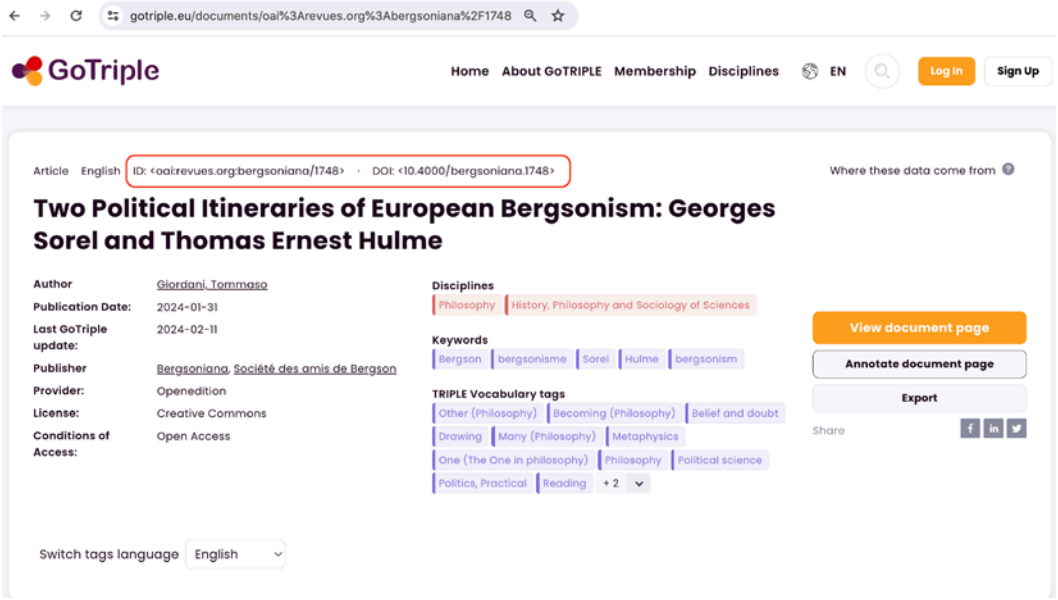

Figure 2. GoTriple document page. The original identifiers of the resource are maintained and shown.

Finally, F4 requires that a searchable index is required to easily access the resources. 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.

Table 2. The FAIR Guiding Principles: Accessibility.

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

All GoTriple data is accessible via the open and standard HTTP protocol (principles A1. and A1.1), providing interfaces both for human (the web site GoTriple.eu) and software programs, the latter in the form of REST APIs, accessible publicly or only after a previous authentication (A1.2). APIs' documentation is available through a link in the GoTriple web site [10].

GoTriple's documents metadata are also accessible for harvesting through the standard OAI-PMH [11] protocol.

 doi:10.20944/preprints202405.0748.v1

Finally, GoTriple copies in its local indexes 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, is no longer accessible (A2.).

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: being 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 [12], 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, the use of shared vocabularies and providing references to other resources. Its guiding principles are presented in Table 3.

Table 3. The FAIR Guiding Principles: 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

In the course of time the GoTriple data model has been formally defined through the TRIPLE Ontology [9] (see principle I1.). This allowed not only to provide a semantic web-compliant, machine understandable description of the three main asset types of the platform, but also of the controlled vocabularies used for specifying the values of the "License", "Condition of Access", "Document Type" and "Disciplines" 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 formalization of the SSH domain in 27 fields of study. This formalization 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 the TRIPLE Vocabulary [16]. The latter consists of a rich controlled vocabulary of over 3.370 linked data SSH-related concepts: it is another significant outcome of the TRIPLE project, which has been formalized in SKOS and provides translations in 11 languages.

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

 doi:10.20944/preprints202405.0748.v1

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. The FAIR Guiding Principles: 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 entities description 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, whose metadata are imported by using the OAI-PMH harvesting protocol with metadata described in Dublin Core.

License and conditions of access for using the imported metadata are preserved and republished on GoTriple by using a controlled vocabulary (see section 2.4), of course if they have been 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 to 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 recognized standards. It is also worth mentioning the possibility for users to easily export document's metadata 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 to the FAIR principles has been described. Beyond the mere declaration of commitment, it is useful to see how all the choices made and presented above translate in practice. Most of these measures proved to be extremely useful and beneficial for the research community at large, as they enabled extended data interoperability and reuse in several contexts.

The first example is provided by VERA [17], the OPERAS' service for participatory research. It is a web-based platform that enables laypersons and researchers to collaborate on SSH-related initiatives of citizen science. Projects in VERA are defined by a set of descriptive attributes, including "Academic subjects". For this, GoTriple's Disciplines have been used. Also, a VERA project can be automatically published on GoTriple and the use of this shared vocabulary 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 a research on identifying similar illustrations in scientific publications [18]. They collected via APIs over 5.600 scientific papers for which their full text in PDF was available. This enabled the team to extract almost 65.000 images that helped them to develop and fine-tune the software methodology at the core of their research.

Finally, through a collaboration amongst the University of Pisa and the Italian company Net7 [19], an Artificial Intelligence driven chatbot was implemented on GoTriple's content. In a similar

 doi:10.20944/preprints202405.0748.v1

way, the full text of over 1,000 articles of a specific SSH discipline was 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 the full compliance with FAIR principles, not all GoTriple's design and, especially, implementation choices have proved to be valid, the worst of them being the implementation of Persistent Identifiers (PIDs). As said, this consists of the URLs of the presentation page of an entity obtained by merging the URL of GoTriple with the original identifier of the content at the remote source, normally those returned as the first one in the harvesting phase, e.g.:

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

For example the PID of a GoTriple document is:

https://www.gotriple.eu/documents/oai%3Arevues.org%3Abergsoniana%2F1748

In general using URL for IDs is a bad idea, as it is proven that over a long time a significant percentage of URLs (over 50%) become inaccessible [20]. Also, linking so strictly an external identifier to GoTriple's ID might be risky. In fact, while harvesting sources it is common to receive multiple IDs for the same resource and the first one that is taken is not always necessarily the one designed to be persistent. It might be an internal ID that can change over time, providing that, if present, the persistent ID (e.g. a DOI) is maintained.

Unfortunately, this consideration arrived too late in the implementation of GoTriple, when already millions of contents had been ingested and published. Fixing this problem, by creating PIDs with a more solid logic for the new but also for the existing content, is one of the goals for the next plan of implementation of GoTriple. One interesting approach to consider is the one used by OpenAIRE for the entities of its graph [21]. Here the PIDs are created by merging an identifier of the source and a hash code generated by processing the remote identifier, by privileging for this operation those known to be persistent (DOI, handle.net, ISBN). Of course, this operation in GoTriple must be planned with great care in order to maintain backward compatibility with the current platform's main identifiers.

Finally, it must also be noted that most of the care in designing and implementing GoTriple's FAIR support has been devoted to documents, which represent the majority and possibly the most valuable content of the Discovery Platform. While the TRIPLE Ontology has recently been expanded to integrate support for projects and profiles, some other FAIR important features, e.g. the possibility to harvest content via OAI-PMH, are at present only available for documents.

4. Conclusions

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

This effort is being pursued under the ATRIUM research project [22], 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) will 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]).

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

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

 doi:10.20944/preprints202405.0748.v1

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data 3, 160018 (2016). doi: 10.1038/sdata.2016.18
2. European Commission, Directorate-General for Research and Innovation, Aronsen, J., Beyan, O., Harrower, N. et al. (2021) *Recommendations on FAIR metrics for EOSC*. Publications Office. doi: 10.2777/70791.
3. Scrum.org web site. Available online: https://www.scrum.org (accessed on 01 April 2024).
4. FitSM web site. Available online: https://www.fitsm.eu/ (accessed on 01 April 2024).
5. FAIR Data Maturity Model Working Group, *FAIR Data Maturity Model. Specification and Guidelines*. Zenodo, 2020. doi: 10.15497/rda00050.
6. TRIPLE Project web site. Available online: https://project.gotriple.eu/ (accessed on 01 April 2024).
7. OPERAS web site. Available online: https://operas-eu.org/ (accessed on 01 April 2024).
8. Deutsche Nationalbibliothek (DNB) web site. Available online: https://www.dnb.de/ (accessed on 01 April 2024).
9. TRIPLE Ontology. Available online: https://www.gotriple.eu/ontology/triple (accessed on 01 April 2024).
10. De Santis, L. (2022). *TRIPLE Deliverable: D6.6 API's Development -RP3*. Zenodo. doi: 10.5281/zenodo.7371832
11. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Available online: https://www.openarchives.org/pmh/ (accessed on 01 April 2024).
12. eTranslation. Available online: https://commission.europa.eu/resources-partners/etranslation_en (accessed on 01 April 2024).
13. COAR Controlled Vocabularies for Repositories: Resource Types 3.1. Available online: https://vocabularies.coar-repositories.org/resource_types/ (accessed on 01 April 2024).
14. SKOS - Simple Knowledge Organization System. Available online: https://www.w3.org/2009/08/skos-reference/skos.html (accessed on 01 April 2024).
15. MORESS - Mapping of research in european social sciences and humanities. Available on line: https://cordis.europa.eu/project/id/HPSE-CT-2002-60060 (accessed on 01 April 2024).
16. TRIPLE Vocabulary. Available online: https://www.semantics.gr/authorities/vocabularies/SSH-LCSH/vocabulary-entries?language=en (accessed on 01 April 2024).
17. VERA - Virtual Ecosystem for Research Activation. Available online: https://vera.operas-eu.org/ (accessed on 01 April 2024).
18. Leszczuk, M., & Dziula, P. (2024). I*mplementation of Software for Searching Similar Illustrations in Scientific Publications.* Opening collaboration for community-driven scholarly communication (OPERAS2024), Zadar, Croatia. Zenodo. doi: 10.5281/zenodo.10958630
19. Bertozzi, A., Abbamonte, M. L., Abaza, A., & De Santis, L. (2024). F*rom Words to Search: GoTriple's AI ChatBot for Efficient Research Engagement*. Opening collaboration for community-driven scholarly communication (OPERAS2024), Zadar, Croatia. Zenodo. doi: 10.5281/zenodo.10977163
20. Stojanovski, J. *PIDs in the SSH - Current state and upcoming challenges*. TRIPLE Booksprint - The role of open metadata in the SSH scholarly communication, Konstancin-Jeziorna (Poland), 7-9 September 2022
21. OpenAIRE Graph Documentation. *PIDs and identifiers*. Available online: https://graph.openaire.eu/docs/data-model/pids-and-identifiers (accessed on 01 April 2024).
22. Advancing fronTier Research In the arts and hUManities (ATRIUM) web site. Available online: https://atrium-research.eu/ (accessed on 01 April 2024).