DOI: 10.5281/zenodo.3937377

Triple

Transforming Research through Innovative
Practices for Linked Interdisciplinary Exploration

[MARCH 31 2020] Advancing Open Scholarship

D1.3 – DATA MANAGEMENT PLAN

Version 1.0 – Draft
PUBLIC

H2020-INFRAEOSC-2019
Grant Agreement 863420

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863420

Disclaimer- "The content of this publication is the sole responsibility of the TRIPLE consortium and can in no way be taken to reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information it contains."

Triple

D1.3. Data Management Plan

Project Acronym:	**TRIPLE**
Project Name:	**Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration**
Grant Agreement No:	**863420**
Start Date:	**1/10/2019**
End Date:	**31/03/2023**
Contributing WP	WP1 Management and Coordination
WP Leader:	CNRS (Huma-Num)
Deliverable identifier	D 1.3
Contractual Delivery Date: 02/2020	**Actual Delivery Date:** 03/2020
Nature: Report	**Version:** 1.0 **Draft**
Dissemination level	**PU**

Revision History

Version	Created/Modifier	Names
0.0	Anas Fahad Khan CNR, Monica Monachini CNR, Laurent Capelli CNRS(HN), Francesca Di Donato Net7, Arnaud Gingold CNRS(OE), Suzanne Dumouchel CNRS(HN), Emilie Blotière CNRS(HN)	First draft including DMPonline template
0.1	Marta Blaszczynska IBL-PAN, Emilie Blotière CNRS(HN), Suzanne Dumouchel CNRS(HN), Francesca Di Donato Net7, Stefano De Paoli Abertay University, Luca De Santis Net7, Christopher Kittel OKMAPS, Simone Kopeinik KC, Peter Kraker OKMPAS.	Completeness of first version
0.2	Maxime Bouillard MEOH, Andrew Pomazanskyi Nuromedia	Peer review

0.3	Maxime Bouillard MEOH, Mélanie Bunel CNRS(HN), Laurent Capelli CNRS(HN), Arnaud Gingold CNRS(OE), Luca De Santis Net7	Second draft including comments of reviewers
1.0	Mélanie Bunel CNRS(HN), Emilie Blotière CNRS(HN), Anas Fahad Khan (CNR)	Final draft version

Abstract

Social Sciences and Humanities (SSH) research is divided across a wide array of disciplines and languages. While this specialization makes it possible to investigate the extensive variety of SSH topics, it also leads to a fragmentation that prevents SSH research from reaching its full potential. Use and reuse of SSH research is low, interdisciplinary collaboration possibilities are often missed, and as a result, the societal impact is limited. TRIPLE, the European discovery solution, addresses these issues: it enables researchers to discover and reuse SSH data, but also other researchers and projects across disciplinary and language boundaries. It provides all necessary means to build interdisciplinary projects and to develop large-scale scientific missions. It will thus increase the economic and societal impacts of SSH resources. Thanks to a consortium of 19 partners, TRIPLE develops a full multilingual and multicultural solution for the appropriation of SSH resources. The TRIPLE platform provides a 360° discovery experience thanks to linked exploration provided by the Isidore search engine developed by CNRS and a coherent solution providing innovative tools to support research (visualisation, annotation, trust building system, crowdfunding, social network and recommender system). TRIPLE imagines new ways to conduct, connect and discover research; it will promote cultural diversity inside Europe; it will support scientific, industrial and societal applications of SSH science; it will connect researchers and projects with other stakeholders: citizens, policy makers, companies, enabling them to take part in research projects or to answer to some of their issues. TRIPLE will be a dedicated service of OPERAS RI and become a strong service in the EOSC marketplace. To conclude, TRIPLE will help SSH research in Europe to gain visibility, to be more efficient and effective, to improve its reuse within SSH and beyond and to dramatically increase its societal impact.

Table of Contents

Table of Figures

Acronyms

CC	Creative Commons
NA	Not Applicable
RI	Research Infrastructure
SSH	Social Sciences and Humanities
TBS	Trust Building System
WP	Work Package

Publishable Summary

This document is made in two main parts: a first part on data summary listing all the produced data by the TRIPLE project and a second part explaining how the TRIPLE project is making FAIR, secured and ethical data.
The known produced data at this stage include 1) interviews data used to co-design the platform 2) collected, enriched and exposed data which are part of the core of the platform 3) innovative functionalities data overlapping the TRIPLE core created by Innovative Services and 4) other contextual data.

To describe the involved data, these are the questions that are addressed for each data in each section following the DMP online template[1] :

- State the purpose of the data collection/generation
- Explain the relation to the objectives of the project
- Specify the types and formats of data generated/collected
- Specify if existing data is being re-used (if any)
- Specify the origin of the data
- State the expected size of the data (if known)
- Outline the data utility: to whom will it be useful

[1] https://dmponline.dcc.ac.uk/

1. DATA SUMMARY

1.1. Platform co-design data

The platform co-design data are related to a mix of social sciences and design research approaches in order to study the TRIPLE platform users and explicit the user needs to co-design core functionalities of TRIPLE. To better understand the working practices of SSH researchers, an initial literature review of existing SSH digital working practices will be conducted to support building of research instruments like interviews and questionnaires.

- State the purpose of the data collection/generation
Data will be collected for the purpose of design and evaluation of a research platform for Social Sciences and Humanities researchers.

- Explain the relation to the objectives of the project
We will collect data in the form of:

- <u>Qualitative interviews</u>

This data is needed for pre-design and identification of user needs and later in the project for evaluation purposes

- <u>Co-design workshops and other focus groups</u>

This data is needed in order to co-design with users some of the core features of the platform, including its governance model

- <u>Questionnaires</u>

One questionnaire will be conducted pre-design for gathering needs and a second one will be conducted post-design for evaluation- questionnaires - other evaluation data => such as A/B testing, walkthrough etc. will be collected for evaluation purposes.

- Specify the types and formats of data generated/collected
Data will come in the following forms:

- <u>Interviews</u>

Both audio recording file and textual transcriptions

- <u>Workshops</u>

Produced material will come in the form sketches and video recording and transcription of the workshops, notes collected by researchers

- <u>Questionnaires</u>

Data will come in the form of tables in CSV format

- <u>Other evaluation data</u>

Data will come in the forms of simple analytics, screen-video recordings, researcher notes

- Specify if existing data is being re-used (if any)
No existing data will be re-used.

- Specify the origin of the data
Interviews, questionnaires, observation metrics. Sources of data are European researchers (the project is Europe wide) mainly, but we are planning a limited number of interviews/workshops also with other stakeholders such as journalists or policy-makers. We will interview these actors and conduct workshops with them. Participants will be selected through contacts of the project partners and via other institutional channels (such as professional mailing lists, social network groups etc.)

- State the expected size of the data (if known)*:* NA

- Outline the data utility: to whom will it be useful
The data will be useful firstly to the project consortium for the design of the platform. They will be used only by all Triple partners and by evaluators under conditions. In the future the data could be reused by researchers interested in Social Studies of Sciences.

1.2. TRIPLE Core data

TRIPLE Core includes a search engine built on ISIDORE technology (provided by Huma-Num). The search engine has to be trained for the different SSH outputs TRIPLE is dealing with, such as data from publications, profiles and projects. For doing so, repositories are trained during the whole process of data enrichment in a multilingual and multidisciplinary context. Three categories of data will be produced: acquisition data, enrichment data and exposition data.

1.2.1. Acquisition Data

- **State the purpose of the data collection/generation**
The purpose is to harvest useful data from external multiple sources in SSH fields (mainly repositories). The TRIPLE platform will only host the harvested metadata and links to the actual resources.

- **Explain the relation to the objectives of the project**
The TRIPLE platform objective is to offer a discovery tool for 3 types of data:

- Research outputs (publications and datasets)
- Profiles
- Projects

The platform will therefore harvest and index any useful resource in the whole SSH research area.

- **Specify the types and formats of data generated/collected**
The collection and indexing process will differ depending on the type of data:

● **Research outputs**: the metadata will be harvested using the OAI-PMH protocol [1] from providers and according to a specific TRIPLE metadata format (XML/oai_dc) combining ISIDORE data model[2] and OpenAIRE guidelines[3]. Existing persistent identifiers (PIDs) like URNs, DOI, ARK or ISBN will be used on the platform. For the publications, in order to ensure consistency of the enrichment process (see "Enrichment data" section), a link to the full text will also be necessary. For the datasets, only the metadata of the set will be harvested and indexed, not the metadata of each unit from the dataset.

● **Profiles:** users [4] will be able to create a TRIPLE account in two different ways by :

- creating a TRIPLE ID directly on the TRIPLE platform including a detailed profile with curriculum vitae (fields of expertise, affiliation etc.…), and to claim publications. Users will also have the possibility to collect their information through their ORCID account, LinkedIn or ResearchGate
- creating a simple account to record requests, follow authors, save files

● **Projects**: the TRIPLE platform will also allow the two possibilities of (a) creation of projects on the platform and (b) of harvesting some from external sources (files, raw data or databases). TRIPLE will harvest the usable projects' information from CORDIS or OpenAIRE and from national funding agencies (ANR in France, for example). Automated interlinking between users' profile (social ID) will make it possible to enrich the projects database.

- Specify if existing data is being re-used (if any)

All the harvested publication metadata will be used for an enrichment process and then converted to RDF to be exposed. It will allow providers and users to collect them in their own environment and to reprocess them. See the Ethics section for more details on GDPR provisions.

- Specify the origin of the data

● Research outputs: in the long term, the origin of the data is any relevant SSH repository in the European research area. In the short term, the first data providers will be the beneficiaries (the partners of TRIPLE Project), and then the members of OPERAS RI.

● Profiles: for the profiles not generated on the platform, they will be collected from ORCID or LinkedIn registries. An authentication and authorization service will provide an identity and access management solution (by EGI check-in[5])

● Projects: they will be mainly harvested from CORDIS, OpenAIRE and National Funders (ANR in France).

- State the expected size of the data (if known)

The size will only be known after the data has been collected. Expectations: 5 million harvested documents minimum (that is ISIDORE current level - index data size is ~200 Go) until 20 million. Depends on how much repositories will be included.

- Outline the data utility: to whom will it be useful

The data collected will be of direct interest to develop the searching and user interaction (annotations, recommendations) features for beneficiaries, the partners of the project.

By the acquisition of data, the platform will be able to provide a wide range of searchable data, profiles and projects for end users.

1.2.2. Enrichment data

- **State the purpose of the data collection/generation**

After the harvesting process, the collected metadata will be enriched using controlled vocabularies to improve their quality and their discoverability by using training machine learning algorithms based on scholarly articles and metadata.

- **Explain the relation to the objectives of the project**

Setting up after the harvesting data, this process of enrichment is closely related to the project because it will allow data to be linked together.

- **Specify the types and formats of data generated/collected**

The enrichment consists of three different actions (Figure 1):

● Classification (or categorization) based on a training scholarly article database and using advanced methods based on statistics and language analysis. Documents are classifying by analyzing their semantic proximity to different categories.

● Normalization using thesauri

● Semantic annotations with disambiguation tool using thesauri and Wikidata database. Data will be added to the TRIPLE metadata schema (XML/oai_dc)

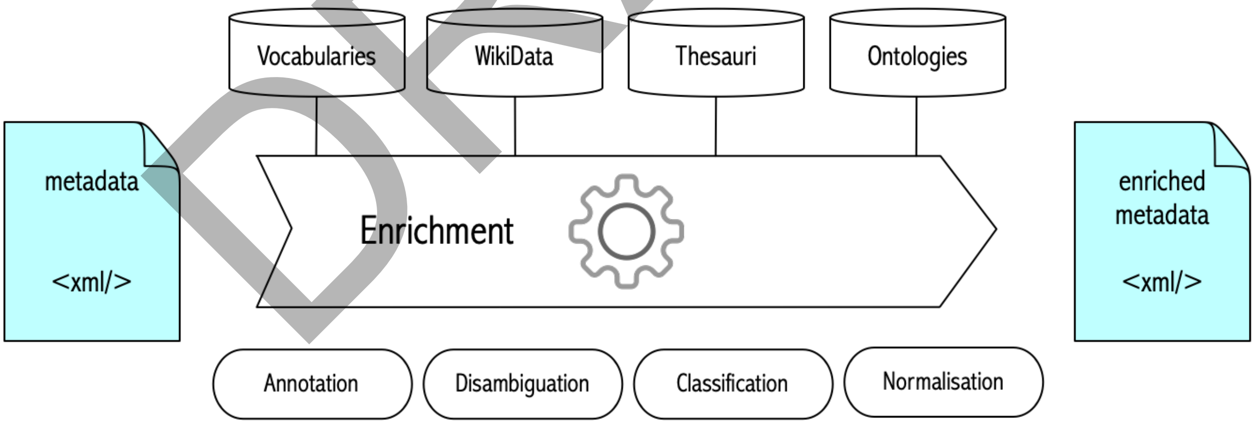

FIGURE 1: Metadata enrichment process

- **Specify if existing data is being re-used (if any)**

The data from enrichment will be reused to continue improving search engine metadata especially for the purpose of machine learning.

- **Specify the origin of the data**

Automatic classifying consists in granting a category for documents using advanced methods based on statistics and language analysis. For categorization, the machine learning model will be trained with scholarly articles from journals referenced in the Directory of Open Access Journals (DOAJ)[6].Categorization, normalization and semantic annotations will be processed by using controlled vocabularies based on existing SSH catalogs[7]. A Multilingual disciplinary Thesauri for the SSH fields in 9 languages will be produced: French, English, Spanish, Portuguese, Italian, German, Polish, Croatian and Greek. Thesauri will be published as Linked Open Data (LOD) datasets in RDF[8] (Resource Description Framework) using Simple Knowledge Organization System (SKOS), a W3C recommendation for the representation of Semantic Web controlled vocabularies. They will also be published as TermBase eXchange (TBX) datasets. Terms thesauri will be aligned to each other.

- **State the expected size of the data (if known)**

The size will only be known after the enrichment process will be launched.

- **Outline the data utility: to whom will it be useful**

Triple users at large for a better researchability experience by the improvement of the quality of data.

1.2.3. Exposition data

- **State the purpose of the data collection/generation**

The purpose is to expose the collected data into the Semantic Web according to the RDF model and to implement a query interface through APIs.

- **Explain the relation to the objectives of the project**

The RDF model allows some interoperability between applications. The query interface allows services to directly look for and retrieve the content indexed in the Triple Discovery platform.

- **Specify the types and formats of data generated/collected**

All the harvested metadata (in TRIPLE format) will be converted into RDF a format and stored in an RDF database (or RDF triplestore [9]) and the *Sparql endpoint* ***[10]*** allows it to be interrogated. The query interface will be provided by ElasticSearch through API. XML data harvested and enriched will be converted and exposed in a JSON representation.

- **Specify if existing data is being re-used (if any)**

Data will be reused to connect with certain Innovative Services.

- **Specify the origin of the data**

Harvested and enriched data at the previous stage.

- **State the expected size of the data (if known)**

The size will only be known after the data has been collected. Expectations: 5 million harvested documents minimum (ISIDORE) = 200 GO until 20 million. Depends on how much repositories will be included.

- **Outline the data utility: to whom will it be useful**

Internal and External → Different core APIs will be created to open TRIPLE data for those who are interested and at least for TRIPLE partners in order to build the User Interface and to connect with the Innovative services.

1.3. Innovative Services Data

The "Innovative Services" are innovative applications and tools that are not part of the core of the TRIPLE platform, but involve data production. These applications and tools will work on top of this core and deliver additional fundamental services for SSH researchers and other TRIPLE stakeholders. Other innovative services would be potentially integrated to this list, but they are still in discussion and not detailed at this stage (a crowdfunding platform or a forum for example).

It includes at this stage:

- A recommender system
- A discovery system
- Open annotation tool
- Trust Building System (TBS)

APIs will make it easy to integrate additional tools as services or data providers in the platform.

1.3.1. User Interaction Data (recommender system)

- **State the purpose of the data collection/generation**

User interaction data will be tracked to form the basis for the project's personalized services. Such services improve the user experience, support decision making and assist in the finding of relevant items and peers.

- **Explain the relation to the objectives of the project**

The data will be used to support the user in finding relevant items (research data, literature, funding opportunities, peers, etc.).

- **Specify the types and formats of data generated/collected**

User interaction will be tracked as event-based data. This means each user event is defined by a specific semantics that encompasses:
- timestamp
- eventid
- userId
- context (this will be specified in accordance with the use case)

- **Specify if existing data is being re-used (if any)**

If applicable for the use case, user profiles (static information about the user) will be reused.

- **Specify the origin of the data**

The data may be collected in all User Interfaces the Triple ecosystem provides.

- **State the expected size of the data (if known)** NA

- **Outline the data utility: to whom will it be useful**

This data will be useful to services that provide user, item or context based personalization, analytical visualizations or any kind of behavioral or social modelling that may serve as a basis for such services. Additionally, end-users will benefit from the improved user experience.

1.3.2. Discovery system

- **State the purpose of the data collection/generation**

To create map representations for the specific purpose of visual presentation of TRIPLE search results.

- **Explain the relation to the objectives of the project**

Central data structure for user interfaces; generated by the TRIPLE/OKMaps machine learning/NLP pipeline which processes raw search result metadata, and generates data structures and file formats required by the user interface

- **Specify the types and formats of data generated/collected**

JSON-files - map representation data

PNG-Images - automatically generated map previews

- **Specify if existing data is being re-used (if any)**

Existing metadata will be re-used which was previously aggregated from different sources/providers.

- **Specify the origin of the data**

TRIPLE core database; search retrieval results; all enriched with map layout and summarization data at generation stage

- **State the expected size of the data (if known)**

On average between 50 KB and 900 KB per individual map representation (uncompressed JSON, metadata only); a few outliers may exist where meta data is especially long (e.g. large abstracts or author lists); the size is also influenced by the result set limit. The total size of the data set is dynamic, as it will be a growing collection of map representations.

- **Outline the data utility: to whom will it be useful**

It will be primarily useful to end users, for whom it will be the technical means to translate their search results into a visual overview

1.3.3. Annotations

- **State the purpose of the data collection/generation**

The purpose is to create digital annotations, i.e., marginalia on digital resources.

- **Explain the relation to the objectives of the project**

Annotations are produced by end users which will use a service directly integrated in the discovery platform.

- **Specify the types and formats of data generated/collected**

Types of data: Highlights, favorites, comments, semantic enrichments, notebooks.
Data formats: RDF, W3C Web annotation data model format serialized as JSON/JSON-LD.

- **Specify if existing data is being re-used (if any)**

The annotation tool (Pundit) will be able to retrieve all existing public annotations carried out on specific web resources, in respect of privacy issues. In this way it will be possible to visualize annotations made before the existence of TRIPLE or by people unfamiliar with the project. The integration with other annotation tools is among the objectives of the project.

- **Specify the origin of the data**

The Pundit Annotation server [11] will allow end users to add annotations directly in the documents.

- **State the expected size of the data (if known)** NA

- **Outline the data utility: to whom will it be useful**

Annotations will be useful for end users to help them to collaborate through documents they will annotate. Annotations will potentially help TRIPLE platform services to collect information (new vocabulary for example) and improve their support.

1.3.4. Trust Building System (TBS)

- **State the purpose of the data collection/generation**

The TBS is built upon the principle of "privacy by design". Therefore, the data collection and generation will be kept to its minimum in order to provide the core functionalities of the system:

- User and group profiles
- Encrypted chat groups
- Newsfeed to publish updates and specific requests
- Featured members to bridge private networks

- **Explain the relation to the objectives of the project**

The TBS is a referral system designed as a new generation of social network informed by collective intelligence techniques, complexity theory, and social sciences. It aims to provide connectivity without sacrificing trust in order to enable multi-stakeholder cooperation.

- **Specify the types and formats of data generated/collected**

Types:

- Authentication account data (user name, email address)
- Profile data (location, company, position, education, about)
- UGC: posts, requests (AES encrypted)
- Chat messages (end-to-end encrypted)

Formats: not answered

- **Specify if existing data is being re-used (if any)**

Potentially, social login.

- **Specify the origin of the data**

ORCID, EGI check-in.

- **State the expected size of the data (if known)**

Not known at this stage.

- **Outline the data utility: to whom will it be useful**

- To users of the TBS
- To users of TRIPLE innovative services?
- To researchers

- To admin team

1.4. Other data, bibliographical data

The TRIPLE project will create contextual data related to its building like working data used by the partners of the project.

- **State the purpose of the data collection/generation**

Bibliographical data will be collected in order to conduct a literature review on existing SSH digital working practices and user research (part of task 3.1 in WP3 about Co-design and user research).

- **Explain the relation to the objectives of the project**

The data collected will be analyzed to inform the building of research instruments (interviews, questionnaire). and will be included in the literature review, part of Deliverable D3.1 (Report on user needs).

- **Specify the types and formats of data generated/collected**

Bibliographic records will be collected and stored in the Zotero open-source bibliographic system and frequently updated by project partners. The data will regularly be exported from Zotero as a CSV file and stored on the TRIPLE project Google drive folder. Available full texts will be downloaded in the PDF format (converted if necessary) and placed in the dedicated folder within the TRIPLE project Google drive.

- **Specify if existing data is being re-used (if any)**

Some of the data, if not generated by the team, is already collected from open sources, thus reused.

- **Specify the origin of the data**

The data will be collected through the method of topical queries on bibliographical databases and by using the snowball method (identifying relevant literature in the bibliographies of key texts).

- **State the expected size of the data (if known)**

The size will be known after all the documents will be collected and will evolve with the regular updating (app. 200 records).

- **Outline the data utility: to whom will it be useful**

It will be useful to all partners within the TRIPLE project (in particular partners involved in WP3 Co-design and user research), and to the OPERAS community more widely, especially those stakeholders interested in user-focused research on digital practices within SSH.

2. FAIR data

2.1 Making data findable, including provisions for metadata

The following types and levels of metadata will be produced and archived:

- **Study-Level Metadata Record.** A summary record in an agnostic format will be created for inclusion in the searchable TRIPLE online catalogue. This record will be indexed with terms from the TRIPLE Thesaurus to enhance data finding.
- **Data Citation with Digital Object Identifier (DOI)**. A standard citation will be provided to facilitate attribution. The DOI provides permanent identification for the data and ensures that they will always be found at the URL specified.
- **Variable-Level Documentation.** TRIPLE will tag variable-level information in the most relevant open standards for SSH i.e. in Data Documentation Initiative (DDI), Text Encoding Initiative (TEI), Metadata Encoding and Transmission Standard (METS), Metadata Object Description Schema (MODS).

Metadata records produced by TRIPLE will be published using the following standard vocabularies: *Component MetaData Infrastructure, Dublin Core Metadata Element Set and DCMI Metadata Terms. Moreover, metadata records published in RDF will use the following linked open data vocabularies: Data Catalog Vocabulary (DCAT), Open Digital Rights Language (ODRL), DDI-RDF Discovery Vocabulary (Disco).*

Annotation data will be published in compliance with W3C Web Annotation standards (data model, vocabulary and protocol).
As regards the specific case of user research data much of the data generated will be qualitative. As such this data will not be made discoverable through automated means. Much of the data will be in the form of textual files or audio/video files.Some metadata will accompany the qualitative data as a text file (.txt) and will be stored with the data at the WP3 leader Secure Storage Drive, where we will report: the name of project, the start/end data, the number of interviews/workshops. Quantitative data will come in the form of CSV files and we will again use a text file with the following metadata: the name of the project, the start/end data, the number of questions, the number of responses. The questionnaire data will be stored with a copy of the questionnaire, for facilitating eventual data reuse. Also, this data will be stored at the WP3 leader Secure Storage Drive.

2.2 Making data openly accessible

TRIPLE data and code will be managed by Huma-Num according to their present standards (see below). Each external service integrated in TRIPLE will be solely responsible for the management of its own data, in full accordance with the TRIPLE guidelines and DMP.

Open Access is the general principle of scientific dissemination in TRIPLE: this means, in practice, that the project grants Open Access to all of the project results, which will be published in Open Access Journals (Gold road) and, when relevant, deposited in Open Access repositories (Green road). All data and metadata will be available in Open Access with open licenses allowing reuse according to Commission requirements. An eventual embargo period will be set-up according to the definition of Plan S[2]. A set of open licenses will be made available to the consortium and for future users of the TRIPLE platform. Hence, the platform will set standards by determining the rules for open research practices and workflows. An effort will take place to set up guidelines to harmonize Open Access and Open Science policies and practices among the various European organizations who will be participating in the platform, in view of developing a shared vision which places Open Access and responsible research at the forefront.

TRIPLE will make research data available to the broader social science and humanities research community. Data will be released under CC open licenses, in order to enable its reuse, and will be deposited in open data repositories such as Zenodo and/or CLARIN/DARIAH and other disciplinary repositories, as well as in the central digital project repository run by CNRS (Huma-Num). Documentation about the software needed to access the data will be included as well as the open source code of the relevant software. The data gathered (research data, publications, profile data, etc.) will be exploited:

- Internally in the form of data enrichment/refinement into innovative platform features (recommender services, visualizations, customized reports etc.) and
- Externally via APIs to 3rd party service providers for further developments.

In the case of public-use data files: These files, in which direct and indirect identifiers have been removed to minimize disclosure risk, may be accessed directly through the TRIPLE repository web site or via well-defined data interfaces from specific tools. After agreeing to Terms of Use, authorized TRIPLE users may download data packages, and unregistered users may purchase the files.

[2] https://www.scienceeurope.org/our-priorities/open-access

2.3 Making data interoperable

We anticipate that the main formats of the data collected and generated will include: *ASCII, tab-delimited (for use with Excel, esp. for survey data), SAS, SPSS, XML (TEI-P5), RDF Serialisation formats (RDF/XML, Turtle, JSON-LD). Documentation will be provided as PDF/A and XML (i.e. TEI, DocBook).*
Metadata records produced by TRIPLE will be published using the following standard vocabularies: *Component MetaData Infrastructure (CMDI), Dublin Core Metadata Element Set and DCMI Metadata Terms. Moreover, metadata records published in RDF will use the following linked open data vocabularies: Data Catalog Vocabulary (DCAT), Open Digital Rights Language (ODRL), DDI-RDF Discovery Vocabulary (Disco).*

2.4 Increase data re-use (through clarifying licenses)

2.4.1 Data License

Open Access is the general principle of scientific dissemination in TRIPLE: this means, in practice, that the project grants Open Access to all project results, which will be published on Open Access Journals (Gold road) and, when relevant, deposited in Open Access repositories (Green road). All data and metadata (**with the exclusion of the User Research Data**) will be available in Open Access with open licenses allowing reuse, according to the Commission requirements**.**

2.4.2. Intellectual Property Rights

All IPR issues will be defined in the Consortium agreement. In any case, IPR will be addressed by taking into consideration the different data types in question. As a general principle, by depositing data within the TRIPLE repository, researchers/authors do not transfer copyright but instead grant permission for TRIPLE to re-disseminate data and to transform data as necessary in order to protect respondent confidentiality, improve usefulness, and facilitate preservation.

3. ALLOCATION OF RESOURCES

3.1 Identify responsibilities for data management in your project

During the project, data management depends on the technical board. The technical board is composed of the four technical WP leaders:

- WP2 - Data acquisition
- WP4 - Integration and building of TRIPLE platform
- WP5 - Development and integration of innovative services
- WP6 -Open Science and EOSC integration

The technical board meets twice a month and invites all members of the consortium to share information. The FAIR Data officer of the European Research Infrastructure, OPERAS [12], will share the responsibility of Data management with IT engineer in charge of the Data management of Isidore [13] (the core of TRIPLE platform).

3.2 Describe costs and potential value of long-term preservation

The FAIRification process is part of the TRIPLE platform development: the data providers will have to provide already FAIRified content, which will be further enriched by TRIPLE to fully ensure its Findability, Accessibility, Interoperability and Reusability. The main related cost of the FAIRification is thus, on one hand, to provide the appropriate support to the data providers, and on the other hand to ensure the effective FAIRness of the data managed by the platform. Therefore, the TRIPLE project allocated specific resources in order to create two specific job positions: a Data Steward for the support to the data providers and a Data Quality Officer for data acquisition and enrichment quality assessment. The Data Steward position will be maintained until the end of the project; the Data Quality Officer position will be maintained during the entire Data acquisition process. The Data Steward and the Data Quality Officer have been hired in February 2020.

4. DATA SECURITY: ADDRESS DATA RECOVERY AS WELL AS SECURE STORAGE AND TRANSFER OF SENSITIVE DATA

Research data from TRIPLE will be deposited within the central digital project repository run by CNRS (Huma-Num) to ensure that the research community has long-term access to the data. The integrated data management plan leverages the capabilities of TRIPLE and its trained archival staff. CNRS (Huma-Num) has a strong expertise in preservation and storage. To avoid the loss of data, CNRS (Huma-Num) is documenting the use of appropriate formats, which are the basis of data interoperability, greatly facilitate the archiving process and making the storage of data independent of the device used to disseminate the data. Different technologies are provided for cold data (i.e. inactive data that are rarely used or accessed), warm data (i.e. data that are analyzed on a fairly frequent basis, but not constantly in play or in motion) and hot data (i.e., data used very frequently and data that administrators perceive to be constantly changing). Last but not least, CNRS (Huma-Num) will provide a long-term preservation service based on the CINES[3] facility (archiving), which is intended for data with a valuable heritage or scientific value.
Selection and Retention: based on this CNRS (Huma-Num) service, TRIPLE will archive full datasets and their related documentation for long term preservation, aiming to support the data through changing technologies, new media, and data formats. TRIPLE will provide a data archive with a long-term track record for preserving and making data available over several generational shifts in technology. Through TRIPLE, CNRS (Huma-Num) will accept responsibility for the long-term preservation of research data upon receipt of a signed deposit form. The collected data will be stored within the existing EU Research Infrastructures. Throughout the life of the project, TRIPLE will ensure that research data is migrated to new formats, platforms, and storage media as required by good practice in the digital preservation community. Good practice for digital preservation requires that an organization address succession planning for digital assets. TRIPLE has a commitment to designating a successor in the unlikely event that such a need arises.

In the specific case of user research data, much of the data will be stored on the WP3 leader Secure Research drive. This will ensure appropriate protection from access (due to encryption being used) as well as recovery in the case of losses (due to back up operated by the University). It may be possible that some user research data will be required to be transferred to third parties:

- <u>The company making transcriptions of audio interviews</u>

We will seek to use a secure Drive for this purpose, which ensures compliance with GDPR and protection of the data.

- <u>Other project partners, for research purposes</u>

[3] https://www.cines.fr

We will seek to use TRIPLE project secure Drive (currently hosted by CNRS) for this purpose, which ensures compliance with GDPR and protection of the data.

5. ETHICAL ASPECTS

Informed consent: For TRIPLE, informed consent statements, if applicable, will not include language that would prohibit the data from being shared with the research community.

The research project will remove any direct identifiers from the data before its deposit within the TRIPLE repository. Once deposited, the data will undergo procedures to protect the confidentiality of individuals whose personal information may be part of archived data. These include: (1) rigorous reviews to assess disclosure risk, (2) modifying data if necessary, to protect confidentiality, (3) limiting access to datasets in which the risk of disclosure remains high, and (4) consultation with data producers to manage disclosure risk. TRIPLE will assign a qualified data manager certified in disclosure risk management to act as steward for the data while it is being processed.

TRIPLE will ensure that personal data processing and management will respect the General Data Protection Regulation (GDPR) provisions, by adopting a privacy by design approach. TRIPLE's privacy policy will be described in a specific document that will be made publicly accessible. Personal data will be collected for the compilation of individual profiles. In this case, data such as first name, surname, encrypted identifiers and IP address will be used to enable the social network functionality which will be part of TRIPLE service. Third-party personal data processing (e.g. interoperable identifiers like ORCID) will depend on their privacy policy. Users will receive clear information when using the service and will be informed of their rights. Other personal data will be automatically collected for the purposes of metrics, especially through the use of cookies. This will enable measurements of site traffic and usage. A privacy policy document will give more details about the duration of personal data storage, but storage for metrics purposes will not exceed 12 months. The responsible for processing in TRIPLE will be the Project Coordination Team (PCT).

[1] https://www.openarchives.org/pmh/

[2] https://isidore.science/sqe. ISIDORE is a French Search Engine in SSH fields developed and maintained by Huma-Num infrastructure.

[3] https://guidelines.openaire.eu/en/latest/

[4] Users or end-users refer to the whole scientific community, citizens, companies, public authorities, policy makers and other socio-economic actors

[5] https://www.egi.eu/

[6] https://doaj.org/

[7] Controlled vocabularies: MORESS categories (categorization) / Lexvo, COAR Resource Type vocabulary, ORCID (normalisation) / GeoNames and TRIPLE SSH vocabulary (dedicated vocabulary for TRIPLE project which is a combination of different SSH catalogs) (annotation-disambiguation)

[8] https://www.w3.org/RDF/

[9] A triplestore disseminates data expressed in the form of "triplets" of information (subject, predicate, object). The Triple Store constitutes the basis of the Data Web (or Semantic Web). The format for modeling and representing these triples is called RDF (Resource Description Framework) and the query language Sparql. These technologies are at the heart of projects such as ISIDORE.

[10] A *Sparql endpoint* is a web interface which allows to query digital information structured in RDF (Resource Description Framework). The interrogation is possible by using the language Sparql, standardized and open language, developed and maintained by the W3C.https://www.w3.org/2009/sparql/wiki/Main_Page

[11] https://thepund.it/

[12] https://operas.hypotheses.org

[13] https://isidore.science