Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration

[23/12/2020]

Advancing Open Scholarship

D1.3 – Data Management Plan

Version 1.1 – Final

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

Deliverable Name

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	**D1.3**
Contractual Delivery Date: 09/2020	**Actual Delivery Date: 12/2020**
Nature: Report	**Version: 1.1 Final**
Dissemination level	**PU**

Revision History

Version	Created/Modifier	Comments
1.0	Anas Fahad Khan CNR,Marta Blaszczynska IBL-PAN, Emilie Blotière CNRS(HN), Maxime Bouillard MEOH, Mélanie Bunel CNRS(HN),Laurent Capelli CNRS(HN), Francesca Di Donato Net7/CNR, Suzanne Dumouchel CNRS(HN), Arnaud Gingold CNRS(OE),Christopher Kittel OKMAPS, Simone Kopeinik KC, Peter Kraker OKMPAS, Monica Monachini CNR, Stefano De Paoli Abertay University, Andrew Pomazanskyi Nuromedia, Luca De Santis Net7	Draft DMP
1.01	Anas Fahad Khan CNR, Marta Blaszczynska IBL-PAN, Emilie Blotière CNRS(HN), Maxime	Deliverable Updated on the basis of DMPOnline

	Bouillard MEOH, Melanie Bunel CNRS (HN), Laurent Capelli CNRS(HN), Francesca Di Donato Net7/CNR, Suzanne Dumouchel CNRS(HN), Arnaud Gingold CNRS(OE), Luca De Santis Net7, Paola Forbes Abertay University, Christopher Kittel OKMAPS, Simone Kopeinik KC, Peter Kraker OKMPAS, Monica Monachini CNR, Panayiota Polydoratou CNRS(OE), Stefano De Paoli Abertay University, Andrew Pomazanskyi Nuromedia, Ondřej Matuška Lexical Computing,Yoann Moranville DARIAH	
1.02	Taina Jääskeläinen (TAU), Alexander König (CLARIN)	Review
1.1	Anas Fahad Khan CNR, Suzanne Dumouchel CNRS(HN), Arnaud Gingold CNRS(OE), Emilie Blotière CNRS(HN), Yoann MORANVILLE DARIAH	Final review

Table of Figures

Acronyms

BM	Business Model
CC	Creative Commons
NA	Not Applicable
RI	Research Infrastructure
SME	Small and Medium Enterprises
SSH	Social Sciences and Humanities
TBS	Trust Building System
WP	Work Package

Table of contents

Glossary

The main references for the glossary are the deliverables D2.1 “Data acquisition plan” and D2.2 “Data harvesting best practices document for data providers”, and Silva C., Ribeiro B. (2010) Background on Text Classification. In: Inductive Inference for Large Scale Text Classification. Studies in Computational Intelligence, vol 255. Springer, Berlin, Heidelberg. https://doi-org.inshs.bib.cnrs.fr/10.1007/978-3-642-04533-2_1

Aggregator: In the context of TRIPLE, an aggregator is an organization that collects, manages, and disseminates the metadata of the scholarly resources’ made available by various providers. The aggregator operates as a standardization body of heterogeneous metadata, either by defining its own requirements, or by relying on existing standards for harvesting and dissemination.

Annotation: Process of adding structure to data (metadata or content) by creating links between a term and elements in a controlled vocabulary. Within the Pundit Annotation service (one of the innovative services of the platform) annotation refers to personal online editing of any web content, such as: highlighting, inserting comments and notes.

Automatic document classification: Refers to a specific categorization process used to enrich documents’ information according to a predefined classification (e.g. a thesaurus). See Categorization.

Categorization: A supervised machine learning classification task, where a training set consisting of documents with previously assigned classes is given as a training set, and a testing set is used to evaluate the models. (Silva C. & Ribeiro B., 2010)

Document: Refers to the information asset related to a specific digital object; it is used to identify single scholarly resources such as publications and datasets.

GOTRIPLE Platform: The public interface where data on SSH scholarly resources, projects and profiles is made available to end-users along with a series of integrated services .

Indexing: The association of a term or terms to a piece of data that serves to facilitate its retrieval.

Normalisation: The translation of specific terms used in a metadata record to a set of pre-defined terms.

Semantic enrichment: A process of adding a layer of topical metadata to content so that machines can make sense of it and build connections from and to it.

TRIPLE Core: The back-end system of the TRIPLE infrastructure that takes care of acquiring, normalizing and semantically enriching data from multiple sources

Introduction

Research carried out in the SSH occurs across a wide array of disciplines and languages. While this specialization makes it possible to investigate a bewildering range of different topics, it also leads to a fragmentation that prevents SSH research from reaching its full potential. Use and reuse of SSH research is not as high as one might desire it to be, interdisciplinary collaboration possibilities are often missed, and as a result, the societal impact of this research can often be limited. TRIPLE strives to address these issues. With a consortium of 19 partners, TRIPLE proposes an integrated multilingual and multicultural solution for the appropriation of SSH resources. The TRIPLE platform will seek to provide an enhanced discovery experience thanks, in large part, to the linked exploration functionalities provided by the ISIDORE search engine, developed and maintained by CNRS-HumaNum[1]. TRIPLE aims to be a coherent solution providing innovative tools to support research (including tools for visualisation, annotation, trust building system, crowdfunding, social network and recommender system). Moreover TRIPLE will propose new ways to conduct and to discover research and will connect researchers, consortiums and institutions with other stakeholders (citizens, policy makers, companies) enabling them to formulate and participate in research projects and respond to other issues. TRIPLE will be a dedicated service of OPERAS RI and seeks to become a strong service in the EOSC marketplace.

When reading this Data Management Plan, one should bear in mind the different types of data and distinct management processes that are herein described. The TRIPLE main data types are the following:

- Co-design research data: As part of WP3, an exhaustive study is being conducted with respect to TRIPLE users and their needs. The data generated by the study has a dedicated data management process (regarding collection, storage, and accessibility). It will be for the most part restricted to WP3 partners. This data is described in Section 1.1.
- TRIPLE data: TRIPLE will primarily collect, process, enrich, and expose metadata from different sources: such metadata is the actual data that the TRIPLE system will manage. This data will be freely searchable on the GOTRIPLE platform, and via the OAI-PMH protocol and a SPARQL endpoint. This data is described in Sections 1.2, 1.3, and 1.4.
- Integrated services data: Another set of data will be generated through TRIPLE's integrated services. Although closely related to the TRIPLE data and platform's activities, this data will be managed independently by each service provider according to specific processes. This data is described in Sections 1.5 and 1.6

[1] https://isidore.science/.

- Publication data: The project will conduct a literature review in relation to its mission and objectives, and produce new publications in this regard. The literature review is related to WP3 and WP6 activities and will be made publically accessible. The publications produced by the TRIPLE consortium will also be publically accessible in appropriate repositories and be published under a CC open license. This data is described in Section 1.7 and 1.8.
- Usage metrics data: Finally GOTRIPLE platform use analytics will be automatically collected, and anonymized where necessary, in order to assess the platform's usage, improve the service, and report on its impact. This data is described in the dedicated Section 3 on Data exploitation.

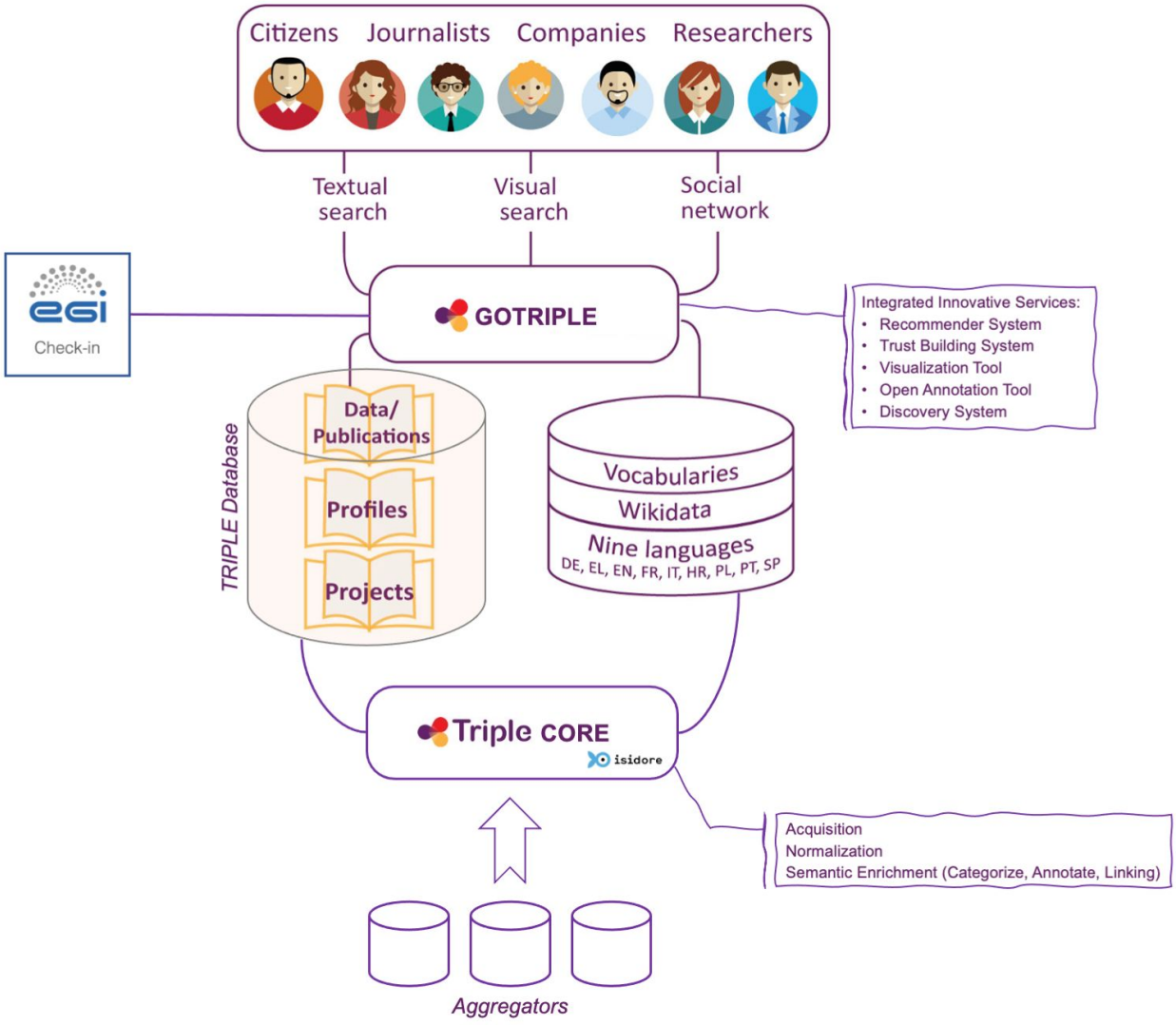

Figure 1 : Representation of GOTRIPLE platform

1 DATA SUMMARY

1.1 Platform Co-design Data

Data Description: *The platform co-design data is related to a mix of social sciences and design research approaches adopted to study the TRIPLE platform users and user needs, essential for co-designing the core functionalities of TRIPLE and for establishing the TRIPLE Forum. In order to better understand the working practices of SSH researchers, an initial literature review of existing SSH digital working practices will also be conducted to support building of research instruments like interviews and questionnaires. Most of the data will be* ***qualitative data****, in the form of interviews or recordings of workshops, focus groups or qualitative evaluation sessions. Most of the material will come from working with SSH researchers and other TRIPLE stakeholders (such as SMEs, or policy makers) seen as potential end-users of the GOTRIPLE platform. Another set of data will come in the form of* ***questionnaire data*** *collected for both the gathering of user needs and evaluation purposes. The data will be collected from SSH researchers (as GOTRIPLE end-users) and other interested stakeholders.* ***Quantitative analytics*** *data will be collected to understand and monitor actual users' behaviour on the platform: this data will include pages visited, actions taken on the platform, events and users' retention.*

State the purpose of the data collection/generation

Data will be collected for the purpose of design and evaluation of the GOTRIPLE research platform - in particular the end-user interface - for SSH researchers. The data will also be used (in anonymised form) for research publications.

Explain the relation to the objectives of the project

For the co-design work, we will collect data through:

- **Qualitative interviews with potential end-users**: This data is needed for pre-design and identification of user needs and later in the project for evaluation purposes. A further set of interviews is needed for establishing the TRIPLE Forum
- **Co-design workshops and other focus groups with potential end-users**: This data is needed in order to co-design with users some of the core features of the platform (with focus on the interface mostly), including its governance model, user profiles, trust system and innovative services.
- **Questionnaires with SSH researchers**: Two questionnaire surveys will be conducted prior to the design for gathering needs (one of these will also be used for the development of the trust building system, see below) and a third one will be conducted post-design for evaluation- questionnaires - other evaluation data, such as A/B testing, walkthrough etc., will also be collected for evaluation purposes.

- **User testing**: Qualitative user evaluation will be done involving users in one-to-one sessions. This activity will be useful to highlight usability problems that will be fixed in iterative sprints.
- **Quantitative metrics:** As the first version of TRIPLE will be published, a metrics analytics system will be integrated to track all user activities on the platform. This data will be used to highlight usability problems and fix them in iterative improvement sprints.

Some metadata will accompany the qualitative data as a text file (.txt) and will be stored with the data, where we will report: *the name of the project (TRIPLE), the start/end of data collection, 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 of data collection, the number of questions, the number of responses*.Raw quantitative data of the platform usage, provided by the metrics analytics service, will not be downloaded but will remain stored on the service servers. These will be used and consulted using the provided graphical user interfaces.

All these types of data are necessary for the achievement of the objectives of WP3.

Specify the types and formats of data generated/collected

Data will come in the following forms:

- **Qualitative Interviews**: Both audio recording file and textual transcriptions. Audio recordings are in .mp3 and .mp4 format, textual transcriptions are .docx format.
- **Workshops**: Materials from these events will come in the form of sketches and video recordings and transcriptions of the workshops, notes collected by researchers; the material from any online whiteboard tool used during the co-design sessions (eg. Miro / Mural). Video recordings are in .mp4 format, textual transcriptions and notes are in .docx format, images from the whiteboards will be in .pdf, .png or .jpg (note that the same type of data will be collected across Tasks 3.2, 3.3, 3.4 and 3.5).
- **Questionnaires**: Data will come in the form of tables in .csv format; Data will come in the form of tables in .docx or Google spreadsheet document.
- **Other evaluation data**: Data will come in the forms of simple analytics, screen-video recordings, researcher notes. Analytics data will be in .csv format, video recording will be in. mp4 format, notes will be in. docx format.
- **Quantitative analytics data**: Data will be stored on the server and database of an external company providing the analytics service. The detailed format of these data is not given. They will be accessible through a web GUI and displayed in form of charts and tables. This is a third party service that complies with the General Data Protection Guidelines (GDPR).

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

No existing data will be re-used. As mentioned above the data may be re-used (in anonymised form) for research publications.

Specify the origin of the data
Sources of data are mostly SSH researchers working in EU Member States (and associated countries), but we are planning a limited number of interviews/workshops, also with other stakeholders such as journalists or policy-makers (also from EU Member States). 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, participation in SSH events etc.)

State the expected size of the data (if known)
Most of this data comes in .mp3 or .mp4 formats). Preliminarily we can say that:

- For Tasks 3.1 and 3.2 together, it should be around 5GB,
- For the research related to Task 3.3 it should be around 3,5GB by the end of the project,
- For Task 3.4 it should be around 10GB (video recording of five workshops of 2,5 hours each),
- For Task 3.5 it should be around 5GB,
- For the User evaluation/testing of Task 3.6, 20 GB (video recording of 20/25 user testing of one hour each).

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. The data will not be shared via public repositories since most of the data is qualitative in nature. Only upon request from a third party the data may be made available (after the request has been evaluated by the data owner). Quantitative data gathered by online analytics services will be used by the designers and developers of the project to improve the usability of the platform. This raw data will not be directly accessed by the consortium, but rather it will be consulted using the service interface. End-users will be informed and will need to provide consent in relation to this data collection. This will be anyway temporary and last only for the period of the evaluation.

1.2 Delivery Platform Raw Data

Data Description: *The delivery platform constitutes the interface between the data providers and the TRIPLE Core. In the present case the data providers are the aggregators and their BUILD Chains (the TRIPLE's primary aggregator and BUILD chain, is ISIDORE). This delivery platform contains raw data dropped by the BUILD chains. The expected raw data is metadata structured according to the TRIPLE data model directly provided by aggregators. These BUILD chains will use a "push" system including the transmission of a signal to drop their data.*

State the purpose of the data collection/generation
The purpose is to obtain raw metadata from BUILD chains (data aggregators) to be injected into the TRIPLE Core in order to set up the enrichment pipeline and create and feed the TRIPLE database. This metadata constitutes the data sources for the TRIPLE Core.

Explain the relation to the objectives of the project
The raw metadata constitutes the *prima materia* of the platform. It is stored in the database after its enrichment.

Specify the types and formats of data generated/collected
Types of Data: Metadata in the TRIPLE metadata format on:

- Documents : Publications and datasets
- Researchers profiles
- Research projects

The metadata may be collected in different file formats, depending on the agreement between the aggregators and the TRIPLE pipeline. We expect the formats to include XML and json.

Specify if existing data is being re-used (if any)
All the raw metadata which has been dropped into a specific file will be used for the enrichment process and then indexed to create the database. See the Ethics section for more details on GDPR provisions.

Specify the origin of the data
Documents: any relevant SSH repository in the European research area collected then harvested by aggregators, research infrastructures.
Profiles: For the profiles not generated on the platform, they will be collected from ORCID or professional registries.
Projects: They will be mainly harvested from CORDIS, aggregated by OpenAIRE and National Funders (e.g. 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 documents minimum (that is ISIDORE's current level - index data size is ~200 GB) to 20 million. This depends on how many repositories will be included.

Outline the data utility: to whom will it be useful
Essential for the creation of the GOTRIPLE platform

1.3 Triple Core Data

Data Description: *TRIPLE Core refers to the core architecture of the platform. It will include: a semantic enrichment pipeline including normalization, annotation and categorization; an indexing process through the search engine ElasticSearch; the TRIPLE database; APIs; connections between the different systems through authentications; the workflow. TRIPLE core data refers to the data which results from the semantic enrichment of the raw metadata described in the last section. A multilingual thesaurus will also be produced as a byproduct.*

State the purpose of the data collection/generation

The raw metadata which has been collected is enriched using controlled vocabularies in order to improve its quality and discoverability using different enrichment processes. The enriched metadata is subsequently indexed into a search engine.

Explain the relation to the objectives of the project

This enriched metadata constitutes one of the main added values of the TRIPLE project. This will improve its quality, one of the objectives of the TRIPLE project. It will be indexed in order to create the TRIPLE database. The RDF model enables interoperability between applications. A query interface will allow services to directly look for and retrieve the content indexed in the GOTRIPLE Discovery platform.

Specify the types and formats of data generated/collected

The semantic enrichment process consists in the enrichment of the raw metadata via the following:

- Automatic document classification (or categorization) based on training a scholarly article database and using advanced methods based on statistics and language analysis. Documents are classified by analyzing their semantic proximity to different categories. These categories are taken from a multilingual thesaurus (described below).
- Normalization using controlled vocabularies
- Semantic annotations with disambiguation tools using thesauri and Wikidata database. Format: XML

In addition we will use ElasticSearch indexing on the resulting data to create json files. The enriched metadata will be converted into RDF and stored in a triplestore.

A multilingual disciplinary thesaurus for the SSH fields in nine languages will be produced. The languages in question are English, French, Spanish, Portuguese, Italian, German, Polish, Croatian and Greek. It will be published as Linked Open Data (LOD) datasets in RDF (Resource Description Framework) using Simple Knowledge Organization System (SKOS), a W3C recommendation for the representation of Semantic Web controlled vocabularies. It will also be published as a TermBase eXchange (TBX) dataset.

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. The indexed data will be used to create the TRIPLE database. Data will be reused to connect with certain Innovative Services through a REST API.

Specify the origin of the data

Raw data combined with enrichment info counts as enriched data which is then indexed. For categorization, the machine learning model will be trained with scholarly articles from journals referenced in the Directory of Open Access Journals (DOAJ)[2] and any other relevant sources. Categorization, normalization and semantic annotations will be carried out using controlled vocabularies based on existing SSH catalogs[3].

State the expected size of the data (if known)

The size will only be known after the data has been collected. Our expectations are as follows: 5 million documents minimum (that is ISIDORE's current level - index data size is ~200 GB) to 20 million. Depends on how many 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 the beneficiaries 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. It will also be useful for TRIPLE users in general for a better research experience via the enhanced quality of the data.

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.

[2] https://doaj.org/

[3] 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

1.4 Machine Learning Data

***Dataset Description**: Machine Learning training data are in the form of abstracts and PDFs in nine different languages in order to classify document abstracts in these languages*

State the purpose of the data collection/generation
Machine learning data is collected in order to train a classifier for classifying document abstracts in nine different languages.

Explain the relation to the objectives of the project
Used to semantically enrich the Delivery Platform Raw Data with annotations.

Specify the types and formats of data generated/collected
Machine learning data will be in the form of multiple XML files.

Specify if existing data is being re-used (if any)
All machine learning data is existing data (largely sourced from the Directory of Open Access Journals (DOAJ) repository) which is re-used for the purpose of machine learning.

Specify the origin of the data
A large quantity of this data has been sourced from the DOAJ repository. Data collection instructions[4] are collected by various different partners of the project (partners provided the title and abstract of their documents, and the full text when available).

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

Outline the data utility: to whom will it be useful
Useful for users of the system as it will be used to enrich metadata records and make resources more findable.
A deliverable under the form of a report on machine learning is due by September 2021 and will give precise details on the salient aspects of this data.

1.5 Innovative Services Data

Dataset Description: *The "Innovative Services'' are applications and tools that are not part of the core of the TRIPLE platform. These applications and tools will work on*

[4] Data collection instructions :https://docs.google.com/document/d/1XTwOoY85WxV_Yb28R58tjLbPe-ttem5D0toDuZp5N-E/edit#heading=h.ip094wx1f28s

top of the GOTRIPLE Platform and deliver additional fundamental services for SSH researchers and other stakeholders. At this stage the list of innovative services comprises[5]: a recommender system; a discovery and visualization system; the Open annotation tool; the Trust Building System (TBS) . Kinds of Innovative Services Data include:[6]

- *User Interaction Data on the GOTRIPLE platform (recommender system)*
- *Visual representations of GOTRIPLE Platform data (Discovery System)*
- *Annotations applied on web resources, in particular web pages and, in the future, PDF documents, applied with the Pundit Open Annotation tool (https://thepund.it)[7].*

1.5.1 Innovative Services Data: User Interaction Data (recommender system)

State the purpose of the data collection/generation
User interaction data on the GOTRIPLE Platform platform will be tracked to form the basis for the project's personalized services. Such services improve user experience, support decision making and assist in the finding of relevant items and peers.

Explain the relation to the objectives of the project
Data will be used to support users in finding relevant items by suggesting them related pieces of information (research data, literature, funding opportunities, peers, etc.).

Specify the types and formats of data generated/collected
User interactions on the GOTRIPLE Platform will be tracked as event-based data. This means that each user event is defined by a specific semantics that encompasses: timestamp; eventid; sessionid; userId (or a coarser identification option, e.g. the IP address); context (this will be specified in accordance with the use case. It might be for example the ID of the resource visited by the user together with its type, the action she performed on the site, etc).

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

[5] *Other innovative services will be potentially integrated with the TRIPLE Core in the future but they are still in discussion and not detailed at this stage (a crowdfunding platform or a forum for example). TRIPLE Core APIs will make it easy to integrate additional tools and services in the platform.*

[6] *At present, the Trust Building System, akin to Pundit, is an external, autonomous system in respect to the TRIPLE Core. No sharing of data amongst these two systems has been envisioned at present: this might change in the course of the project, whether the result of the specific user design research activity (carried on within WP3) will suggest a tighter integration amongst these two systems.*

[7] *In particular Pundit will have a loosely coupled relationship with the TRIPLE Core. No data will be shared between these two systems. The Annotation data is stored in fact in an autonomous fashion by the Pundit service, which can be seen as "external" in relationship with the TRIPLE Core.*

Besides interaction data, if applicable for the use case and especially if available, user profiles (static information about the user) will be reused by the Recommender System.

Specify the origin of the data
These data may be collected from all the User Interfaces provided by TRIPLE ecosystem, in particular from those of the GOTRIPLE Platform.

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 to users, items or contexts based on 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.

1.5.2 Innovative Services Data: Discovery system and visualization

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
The service will expand the User Interface of GOTRIPLE with an advanced interactive search service, enriched with sophisticated visualizations. It will be empowered 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, stored in the TRIPLE Core and which was previously aggregated from different sources/providers, will be re-used for the Discovery System.

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 50KB and 900KB per individual map representation (uncompressed JSON, metadata only); a few outliers may exist where metadata 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 dataset 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.

1.5.3 Innovative Services Data: Annotations

State the purpose of the data collection/generation
To collect digital annotations, i.e., marginalia on digital resources. The purpose of this service is, in fact, to allow users to apply digital annotations on web resources, HTML pages to start with but in perspective on PDF documents as well. In particular the possible specific purposes of a Pundit user (the so-called "motivation" as defined in the W3C Web Annotation standard) are (together with the corresponding Web Annotation definitions):

- Highlighting a fragment of text (**oa:highlighting**)
- Commenting a fragment of text (**oa:commenting**)
- Applying a "semantic annotation" (therefore a formal "statement" described as an RDF "triple") on the whole web resource or on a fragment of its text (**oa:linking**). The new version of Pundit, currently in development, will also allow the association of a free-form text "tag" to an annotation (**oa:tagging**) by using underneath the same mechanism of a semantic annotation
- A social interaction with an existing annotation, as "Like"/"Dislike" or a reply to it (**oa:moderating**).

Explain the relation to the objectives of the project
Annotations are produced by end users who use the Pundit external service, federated in the platform. This federation will allow for example:

- The direct annotation of GOTRIPLE Platform pages
- Opening the external web resources, indexed with the TRIPLE Core, with the Pundit annotation tool automatically activated
- Presenting the number of public annotations that users have applied on resources indexed by the TRIPLE Core.

Specify the types and formats of data generated/collected
Data types: annotations on web resources (highlights, comments, semantic annotations, social interactions, tags); notebooks, which are containers of annotations.
Data formats: internally Pundit stores data in a proprietary format (at present in a RDF Triple Store, in Elasticsearch indexes for the new version currently in development). Users are able to export their annotations in several open formats, including JSON and ODT. Moreover in the new version Pundit will provide an open API that returns annotation data in a fully compliant format with the W3C Web annotation standard and serialized as JSON-LD.

Specify if existing data is being re-used (if any)
All existing public annotations applied with Pundit (even in the past, prior to the beginning of the TRIPLE project) can be visualised by users activating this tool on any web resource, taking of course any privacy issues fully into consideration. Finally, interoperability with other annotation tools (in particular with Hypothes.is) is amongst the goals currently addressed in the development of the new version of Pundit.

Specify the origin of the data
Pundit allows end users to add annotations directly on web documents. Users' annotations are stored in the Pundit infrastructure, which is external from the TRIPLE Core. Moreover, Pundit will reuse profile data returned by authentication services, in particular from EGI AAI Check-In, Google and Facebook.

State the expected size of the data (if known)
The size varies according to the dimension of a single annotation (e.g. the amount of text that is annotated, its type, the complexity of the relationship expressed with it, etc.). On average we can estimate between 1k to 5k of data per annotation. Currently in Pundit around 50,000 annotations are stored.

Outline the data utility: to whom will it be useful
Pundit is a personal utility tool that allows end-users to "take notes" on web resources. Of course annotations are more than simple "textual notes": moreover they are collected in notebooks that can be easily exported in open formats, allowing the user to easily reuse them (e.g. to base a specific research on them). Also annotations might be a useful collaboration strategy for teams that are analysing the same web document.

1.5.4 Innovative Services Data: 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: Proprietary
Workshop and Questionnaire data: same format as 'Platform co-design data'

Specify if existing data is being re-used (if any)
Data from ORCID will be re-used.

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 the users of the TBS; possibly to the users of TRIPLE and its innovative services; to researchers; to the administrative team.

1.6 Authentication Service EGI Check-In

Dataset Description*: For user authentication, the integration of the EGI AAI Check-in service has been envisioned in TRIPLE. This way the GOTRIPLE Platform (and its Innovative Services) will not have to manage explicit authentication credentials, but simply act as "Service Providers" in an authentication scenario in which EGI AAI Check-in acts as the Identity Provider. EGI AAI Check-in is integrated with plenty of existing authentication services, including those of the universities and other main research centres in Europe and of the most common social networks. The integration with EGI AAI Check-in allows to fetch some of the user's credentials, the so-called "Claim Values" as described in the official EGI documentation[8]. In details, all the currently non-deprecated claims are:*

- ***openid.sub****: An identifier for the user, unique among all EGI accounts and never reused.*
- ***name****: The user's full name, in a displayable form (e.g. John Doe)*
- ***given_name****: The user's first name (e.g. John)*
- ***family_name****: The user's last name (e.g. Doe)*

[8] https://wiki.egi.eu/wiki/AAI_guide_for_SPs#Claims

- ***preferred_username****: The username by which the user wishes to be referred to (e.g. jdoe)*
- ***email****: The user's email address (e.g. john.doe@example.org)*
- ***eduperson_scoped_affiliation****: The user's affiliation within a particular security domain/scope (e.g. member@example.org)*
- ***eduperson_entitlement****: The user's entitlements expressed as group/VO membership/role information, defined in the specific EGI AAI Check-In syntax.*

As mentioned above, further claims are currently returned by the system (e.g. acr that identifies the "Level of Assurance/LoA" of the actual identity provider the user has used for authentication), but, according to the official EGI documentation, they are deprecated and therefore they might disappear from the service's response in the future.

State the purpose of the data collection/generation

As specified above, EGI AAI Check-In can provide a set of data regarding the user that has been authenticated with this service. In TRIPLE this data will be used for: identifying the user (e.g. openid, email), for personalising the user interface once she has logged in (name, given_name, family_name, preferred_username), to communicate with her (email) or to identify her affiliation and rights (eduperson_scoped_affiliation, eduperson_entitlement).

Explain the relation to the objectives of the project

In TRIPLE there will be the need to recognize users for some authenticated services (e.g. the Open Annotation tool Pundit and the TBS). EGI AAI Check-In provides a minimal set of user data (see above) that TRIPLE services will exploit for user identification.

Specify the types and formats of data generated/collected

User data returned by EGI AAI Check-in are called "Claims".

Types:

- Identification (opeind): sub
- profile: name, given_name, family_name, preferred_username
- email
- eduperson_scoped_affiliation
- eduperson_entitlement

See above and the official EGI AAI Check-In documentation at:
https://wiki.egi.eu/wiki/AAI_guide_for_SPs#Claims

Formats: EGI Claims are represented by a JSON object that contains a collection of name and value pairs, returned by the EGI Check-in UserInfo Endpoint. The latter is an OAuth 2.0 protected resource that provides specific information about the authenticated End-User.

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

All users authenticating with EGI AAI Check-in are supposed to have an already existing account (e.g. on their research centres' Directory Services, on Social Networks etc.), therefore the EGI integration will apply by default a reuse of user data.

Specify the origin of the data
EGI AAI Check-In service is the origin of user data, for those users authenticated with this service. In fact, even if Check-In acts as a sort of "proxy" by authenticating the user with multiple identity providers (social networks, research centres' directory services,...), an actual user profile is created in EGI.

State the expected size of the data (if known)
For each user, data returned by EGI AAI Check-In (see "Claims" defined above) are very limited in size (approx. between 200 bytes to 1k per user)

Outline the data utility: to whom will it be useful
All TRIPLE components (Core and Innovative Services) that need a user authentication service. At the present day, the TBS and the Annotation Tool Pundit will certainly integrate this EGI service, making use therefore of the data returned by it.

1.7 Bibliographical Data

> ***Dataset Description****: Bibliographical data have been 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). Another literature review has been made in WP6 on the different publications related to the EOSC in order to prepare the GOTRIPLE platform integration. This data comes in the form of publications (papers, books, reports etc.)*

State the purpose of the data collection/generation
Bibliographical data has been collected in order to conduct a literature review on existing SSH digital working practices and user research (part of task 3.1 in WP3). Another literature review has been carried out in WP6 on the different publications related to the EOSC in order to prepare the GOTRIPLE platform integration.

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, was already collected from open sources, thus reused.

Specify the origin of the data
The data has been 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 is approx 300MB

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), and to the OPERAS community more widely, especially those stakeholders interested in user-focused research on digital practices within SSH.

1.8 Communication Data

> ***Dataset Description****: publications made on the TRIPLE project (powerpoints, abstracts, papers, workshops, webinars, posters and scientific articles). This data is public and accessible via a DOI.*

State the purpose of the data collection/generation
This data is generated under the form of public presentations within the Work Package 8 (Communication and Dissemination) to effectively communicate the project and its potential benefits to stakeholders, to communicate research findings, to stimulate an ongoing interest in the work of the project, and to build an overall awareness of the project and its goals.

Explain the relation to the objectives of the project
Publications are disseminated in different ways and to SSH researchers' communities to inform about the objectives and ongoing of the project. Presentations and posters have been created for attending several European conferences dedicated to digital humanities, such as for instance the Open Science Conference 2020, ICTeSSH conference 2020 and EGI conference 2020. Workshops have also been organized during DH European events such as the LREC2020 and the EOSC-hub/FREYA/SSHOC webinars to discuss eventual synergies with

other European projects. All those presentations constitute communication data that is stored in a Google drive folder and in a secured storage interface, harvested by the coordinator (sharedocs).

Specify the types and formats of data generated/collected
We deal here with interoperable formats in Microsoft Office (.doc, .xls, .ppt) and in Open Office (.odt, .ods, .png). We also use open formats for pictures (.jpeg, .gif and .png) as it includes powerpoints, published papers, posters and scientific articles. The most commonly used formats are .ppt, .pdf, and .docx and required formats (HTML) for publications in scientific publishers such as MDPI, a publisher of Open Access Journals.

Specify if existing data is being re-used (if any)
This data is reused internally and is public so it is reusable externally. It is accessible either in Zenodo platform via a DOI or via a permanent url link.

Specify the origin of the data
This data is generated mostly by Work Package 8 but any partner attending an event or invited to take part in a webinar or conference, is likely to create communication data in close collaboration with the WP8 leader in order to maintain the project communication standards (logo and style guidelines) and language elements related to the project and defined by the WP8. Guidelines and toolkit are accessible within the consortium to ensure harmonized communication data.

State the expected size of the data (if known)
The size is not known but will remain minimal (few GBs)

Outline the data utility: to whom will it be useful
Communication data is part of the external communication strategy. This data is useful to disseminate the results of the project and useful to the whole consortium and also to the broader community of SSH researchers in order to learn about the platform and its functionalities to facilitate and enhance the value of their work.

2. FAIR DATA

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)
TRIPLE Core data consists of enriched versions of raw metadata obtained from aggregators and their BUILD chains. This raw metadata describes documents, projects and researcher

profiles. The enriched metadata will be utilised for the purposes of data discovery and used by innovative services. It will also be made available in the ways described above.
As regards the specific case of co-design research data much of the data generated will be qualitative. As such, this data will not be made discoverable through automated means. Each project partner responsible for the user research data collection in WP3 will store the data on their own server (or chosen GDPR compliant service). Questionnaire data will be stored with a copy of the questionnaire, for facilitating eventual data reuse, in the WP3 leader's Secure Storage Drive.

Outline the identifiability of data and refer to standard identification mechanisms. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
Identifiers are mandatory for the raw metadata collected from the aggregators. As far as possible aggregators' identifiers (DOI, Handle, ark and local identifier) for documents, research projects and researchers' profiles will be re-used when possible. In the case of absence of identifiers, the TRIPLE system will be able to generate reusable PIDs.

Outline naming conventions used
File naming conventions for TRIPLE data relate only to the internal processes of the TRIPLE system and will not be disseminated. With regard to Co-design research data, the file naming practice for recorded interviews was to use the initials of the interviewee (or their pseudonym) followed by the number of the interview, the project name and then the date when the interview was carried out.

Outline the approach for clear versioning
For the current ISIDORE platform, there is an automatic update of the documents in the form of a monthly programmed replenishment. The intention is to maintain and expand this provision for the TRIPLE information system.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how
The raw data delivered by aggregators will be conform to the TRIPLE data model built by using the schema.org[9] vocabulary. Any annotation data to be published will be published in compliance with W3C Web Annotation standards (data model, vocabulary and protocol).

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so
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 and, when relevant, deposited in Open Access repositories. All data and metadata will be available in Open Access with open licenses

[9] https://schema.org/. See D.1 "Data Acquisition Plan" for more details.

allowing reuse according to European Commission requirements. An eventual embargo period will be set-up according to the definition of Plan S[10]. 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. Public-use data files will have direct and indirect identifiers removed to minimize disclosure risk.
Here are the accessibility policies for all the other data types:

- Co-design research data: for the most part restricted to the WP3 partners.
- TRIPLE data: freely accessible on the GOTRIPLE platform, and via OAI-PMH protocol and a SPARQL endpoint.
- Integrated services data: the integrated services internal data will not be shared with the TRIPLE information system and will therefore not be publicly accessible.
- Usage metrics data: this data will not be made public except in the form of fully anonymized quantitative figures.

Specify how the data will be made available
TRIPLE data will be to the broader social science and humanities research community under a CC open license, in order to enable its reuse. The TRIPLE data gathered (on datasets, publications, profiles, 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.

As far as the Open Annotation tool (Pundit) is concerned, users can export the annotations stored in their “notebooks” in open formats, including JSON and ODT. This feature is already available in the current version of the service. Moreover, the new version, currently in development, will also provide a search API that returns annotations as JSON-LD data formatted according to the specifications of the W3C Web Annotation standard.
As stated above, TRIPLE consortium publications will be deposited in open data repositories, such as Zenodo or in the central digital project repository run by CNRS (Huma-Num).

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?
Documentation about the software needed to access the data will be included as well as the open source code of the relevant software. The selection of the repository for such documentation is on-going.

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

Specify where the data and associated metadata, documentation and code are deposited
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.

Specify how access will be provided in case there are any restrictions
In the case of the data with restricted access listed above, it is not planned at the moment to provide broader access.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.
Regarding TRIPLE data, 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: relevant Component MetaData Infrastructure (CMDI) profiles, 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). Annotations data serialised in JSON-LD according to the W3C Web Annotation standard specifications.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible
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.
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.

2.5 Allocation of resources

Clearly 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 [11], will share the responsibility of Data management with IT engineer in charge of the Data management of ISIDORE[12] (the core of TRIPLE platform).

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 hire the following in February 2020: a Data Steward for the support to the data providers and a Data Quality Officer for data acquisition and enrichment quality assessment.

3. Data exploitation

Usage metrics will be collected on the GOTRIPLE platform as soon as the first beta version is ready (June 2021), in order to be able to provide those data to WP7 in relation to its task on the sustainability of the platform. These anonymized usage metrics will provide useful information on the platform's usability, dynamics, and potential enhancements; the usage metrics will be exploited in the Plan for Exploitation and Dissemination of Results (PEDR). This data will remain private, even if an analysis of this data may be published on the platform to provide insights to users and other stakeholders. The usage metrics will be quantitative data similar to the quantitative data generated for WP3, but they will be generated only after the creation of GOTRIPLE, or at least after the beta version is ready. The usage metrics will have an important role in ensuring the sustainability and soundness of GOTRIPLE's business model.

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

[12] https://isidore.science

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Relevant data from TRIPLE will be deposited within a central digital project repository run by CNRS (Huma-Num) to ensure that the research community has long-term access to the data. We plan to leverage 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) will make use of appropriate formats in order to ensure data interoperability, facilitate the archiving process and making the storage of data independent of the device used to disseminate the data. CNRS (Huma-Num) will provide a long-term preservation service based on the CINES[13]3 facility (archiving), which is intended for data with a valuable heritage or scientific value.

Throughout the life of the project, TRIPLE will ensure that its 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 addresses succession planning for digital assets. To this end TRIPLE is committed to designating a successor in the unlikely event that such a need arises.

Here are the secure storage provisions for the other sources of data:

- Co-design research data: much of the data will be stored on the WP3 leader Secure Research drive (this includes data from tasks 3.1, 3.2 and most of 3.6). This will ensure appropriate protection from unauthorised access (due to encryption being used) as well as recovery in the case of losses (due to back up operated by the University).
- Trust Building System data (Task 3.3): all of the data is stored on Google cloud/Drive.
- Data collected by Net7 in Tasks 3.4 and 3.6: it will be stored in an offline secure data storage.
- The data collected by the partner OKMaps during their workshop will be processed and stored by Open Knowledge Maps and TRIPLE.
- The data collected by EKT for Task 3.5 will be stored in EKT' s offline multimedia server. In this case, for processing and storage, third parties services that comply with the General Data Protection Guidelines (GDPR), such as Google Drive, may be used. All data will be anonymized after completion of the workshop.

It may be possible that some of the Co-design research data will be required to be transferred to third parties:

1. The company making transcriptions of audio interviews
2. The company offering the analytics service used for the evaluation
3. Other project partners, for research purposes

In the case of 1, we will seek to use a secure Drive for this purpose, which ensures compliance with GDPR and protection of the data. In the case of 2 the service is GDPR

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

compliant. In the case of 3. 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

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Informed consent: For TRIPLE, informed consent statements, if applicable, will not include content that would prevent 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).