10.5281/zenodo.7835582

Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration

[31/03/2023]

Advancing Open Scholarship

D1.3 – Data Management Plan

Version 3.0 – 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."

This deliverable is licensed under a Creative Commons Attribution 4.0 International Licence

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	**D1.3**
Contractual Delivery Date: 03/2023	**31 March 2023 Update**
Nature: Report	**Version: 3.0 Update**
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 OKMAPST, 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 Bouillard MEOH,	Deliverable Updated on the basis of DMP Online

	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, Paula 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	Update
2.0	Emilie Blotiere CNRS (HN), Arnaud Gingold CNRS(OE), Stefanie Pohle (MWS), Francesca di Donato (CNR), Lottie Provost (CNR)	Final review
2.1	Emilie Blotiere CNRS (HN), Francesca Di Donato (CNR), Lottie Provost (CNR), Erzsébet Toth-Czifrà (DARIAH), Luca De Santis (Net7), Giulio Andreini (Net7), Arnaud Gingold (OE), Stefano De Paoli (AU), Haris Georgiadis (EKT)	Update (Dec. 31 2021)
2.2	Francesca Di Donato (CNR), Lottie Provost (CNR), Luca De Santis (Net7), Arnaud Gingold (OE), Maciej Maryl (IBL), Christopher Kittel (OKMAPS)	Update (July 31 2022)
2.3	Francesca Di Donato (CNR), Lottie Provost (CNR), Emilie Blotière CNRS (HN), Luca De Santis (Net7), Arnaud Gingold (OE), Stefano De Paoli (AU), Haris Georgiadis (EKT), Sona Arasteh (MWS), Iraklis Katsaloulis (EKT).	Update (November 30 2022)
3.0	Arnaud Gingold (OE), Lottie Provost (CNR), Emilie Blotière CNRS (HN), Francesca Di Donato (CNR), Luca De Santis (Net7), Stefano De Paoli (AU), Sona Arasteh (MWS), Iraklis Katsaloulis (EKT), Marta Blaszczynska (IBL-PAN) Julien Homo (Foxcub), Simone Kopeinik (KC), Peter Kraker (OKMPAS),Gaël van Weyenberg (MEOH).	Update (March 31 2023)

Table of Figures

Acronyms

BM	Business Model
CC	Creative Commons
NA	Not Applicable
LOD	Linked Open Dataset
RDF	Resource Description Framework
RI	Research Infrastructure
SCRE	Semantic Content Retrieval Engine
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 organisation that collects, manages, and disseminates the metadata of the scholarly resources’ made available by various providers. The aggregator operates as a standardisation 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 GoTriple platform) annotation refers to personal online editing of any web content, such as: highlighting, inserting comments, tags, semantic assertions 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 of documents with previously assigned classes to create the TRIPLE classification model

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, normalising and semantically enriching data from multiple sources. This component in TRIPLE is called SCRE (Semantic Content Retrieval Engine).

TRIPLE data providers: In the context of TRIPLE, a Provider is an organisation that manages, collects, and disseminates scholarly resources. It operates as the manager of one or various data repositories, archives, or publishing platforms that GoTriple harvests. A Provider enriches the data it is responsible for with minimal metadata facilitating its dissemination, and acts as the primary dissemination body of the data and its metadata

Introduction

Research carried out in the SSH occurs across a wide array of disciplines and languages. While this specialisation 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 GoTriple platform will seek to provide an enhanced discovery experience with linked exploration functionalities by leveraging the experience of the ISIDORE search engine, developed and maintained by CNRS Huma-Num[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 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. GoTriple is 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 was conducted with respect to TRIPLE users and their needs. The data generated by the study had a dedicated data management process (regarding collection, storage, and accessibility). Data is restricted to WP3 partners. This data is described in Section 1.1.
- TRIPLE data: TRIPLE primarily collects, processes, enriches, and exposes metadata from different sources: such metadata is the actual data that the TRIPLE system is managing. This data is freely searchable on the GoTriple platform, via dedicated APIs. An OAI-PMH endpoint has been also developed to ease the integration with selected partners. This data is described in Sections 1.2, 1.3, and 1.4.
- Integrated services data: Another set of data has been generated through TRIPLE's integrated services. Although closely related to the TRIPLE data and the GoTriple

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

platform activities, this data is managed independently by each service provider according to specific processes. This data is described in Sections 1.5 and 1.6

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

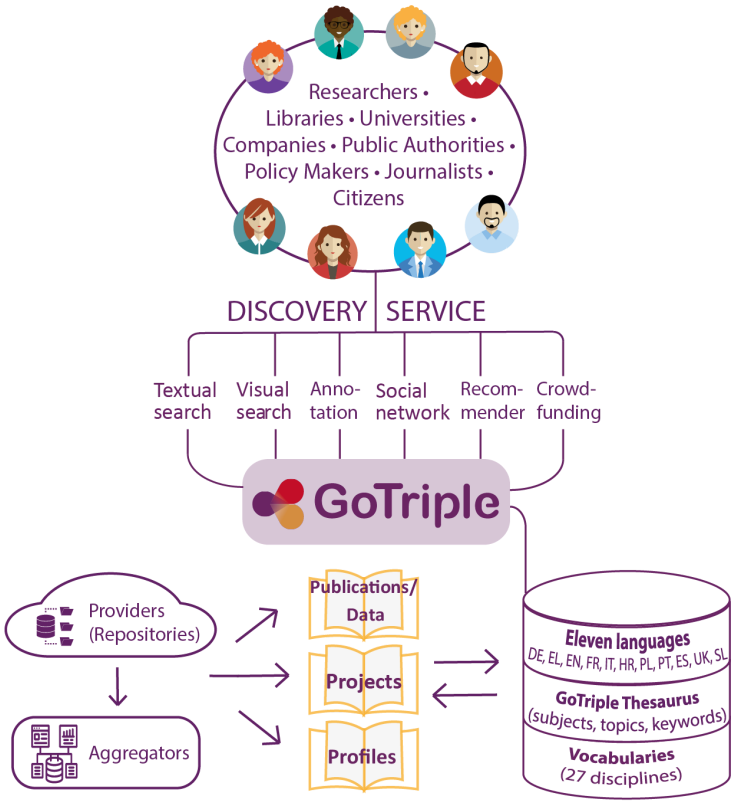

FIG.1 : Representation of GoTriple platform

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 GoTriple platform users and user needs, essential for co-designing the core functionalities of TRIPLE and for establishing the GoTriple beta testers mailing list. In order to better understand the working practices of SSH researchers, an initial literature review of existing SSH digital working practices was conducted to support building of research instruments like interviews and questionnaires. Most of the data is* ***qualitative data****, in the form of interviews or recordings of workshops, focus groups or qualitative evaluation sessions. Most of the material came 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 came in the form of* ***questionnaire data*** *collected for both the gathering of user needs and evaluation purposes. The data was collected from SSH researchers (as GoTriple end users) and other interested stakeholders.* ***Quantitative analytics*** *data is collected to understand and monitor actual users' behaviour on the platform: this data includes pages visited, actions taken on the platform, events and users' retention.*

State the purpose of the data collection/generation

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

Explain the relation to the objectives of the project

For the co-design work, the data was collected through:

- **Qualitative interviews with potential end users**: This data was needed for pre-design and identification of user needs and later in the project for evaluation purposes. A set of interviews/focus groups were conducted for Task 3.5, which are relevant to the needs of the TRIPLE Forum and platform governance.
- **Co-design workshops and other focus groups with potential end users**: This data was 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, dashboard, the trust building system and innovative services.
- **Questionnaires with SSH researchers**: Three questionnaire surveys were conducted prior to the design for gathering needs (one of these was also used for the development of the trust building system) and a fourth one is being conducted post-design for evaluation.

- **User testing**: Qualitative user evaluation was carried out involving users in one-to-one sessions. This activity was relevant to highlight usability problems which are fixed in iterative sprints.
- **Quantitative metrics:** As the first version of the GoTriple platform was published, three metrics analytics systems were integrated to track all user activities on the platform. This data is used to highlight usability problems and fix them in iterative improvement sprints.

Some metadata accompanies the qualitative data as a text file (.txt) and is stored with the data, where the following information is reported: *the name of the project (TRIPLE), the start/end of data collection, the number of interviews/workshops*. Quantitative data comes in the form of .csv files and a text file with the following metadata is used: *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, has not been downloaded but remains stored on the service servers. These are 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 comes 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 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 are in .pdf, .png or .jpg (note that the same type of data was collected across Tasks 3.2, 3.3, 3.4 and 3.5).
- **Questionnaires**: Data comes in the form of tables in .csv format; Data comes in the form of tables in .docx or Google spreadsheet documents.
- **Other evaluation data**: Data comes in the forms of simple analytics, screen-video recordings, researcher notes. Analytics data is in .csv format, video recording is in. mp4 format, notes are in. docx format.
- **Quantitative analytics data**: Data is stored on the servers and databases of the external companies providing the analytics services. The detailed format of these data is not given. They are accessible through web GUIs and displayed in the form of charts and tables. These are third party services that comply with the General Data Protection Guidelines (GDPR).

Specify if existing data is being re-used (if any)
No existing data has been re-used.

Specify the origin of the data
Sources of data are mostly SSH researchers working in EU Member States (and associated countries), but a limited number of interviews/workshops with other stakeholders such as journalists or policy-makers (also from EU Member States) have also taken place. These actors have been interviewed and workshops were conducted with them. Participants were 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. The size of the data is indicated below

- For Tasks 3.1 and 3.2 together, it is around 9GB,
- For the research related to Task 3.3 it is around 3,5GB,
- For Task 3.4 it is around 8GB (video recording of five workshops of 2,5 hours each),
- For Task 3.5 it is around 3GB,
- For the User evaluation/testing of Task 3.6, 20 GB (video recording of 26 user testing of one hour each).

Outline the data utility: to whom will it be useful
The data was useful firstly to the project consortium for the design of the platform. The data is not shared via public repositories since most of the data is qualitative in nature. Only upon request from a third party may the data be made available (after the request has been evaluated by the data owner). Quantitative data gathered by online analytics services is used by the designers and developers of the project to improve the usability of the platform, and will be used also after the end of the funding period. This raw data is not directly accessed by the consortium, but rather it is consulted using the service interface. End users are informed and need to provide consent in relation to this data collection. This will last beyond the funding period as a way to monitor the health of the platform community.

1.2 TRIPLE Core Data

> **Data Description:** *TRIPLE Core refers to the core architecture of the platform. It consists of: a semantic enrichment pipeline including acquisition of content (publications/datasets, projects, authors) from multiple sources (aggregators and providers alike), normalisation, annotation and classification; an indexing process through the search engine ElasticSearch; the TRIPLE database; APIs; connections between the different systems through authentication. TRIPLE core data refers to the data which results from the semantic enrichment of the raw metadata acquired. A multilingual thesaurus, produced by the WP2 team, was also used to perform annotation of content.*

The partner Net7 developed a platform named SCRE to manage the data acquisition and enrichment in GoTriple. Through the core pipeline, metadata regarding publications and projects for the Social Sciences and Humanities are automatically harvested, mapped in the TRIPLE data model, curated, enriched and finally saved in the GoTriple platform's indexes. SCRE imports publications metadata from OAI-PMH endpoints, OpenAIRE and Isidore data dumps. For projects the main data source is the CORDIS database of EU funded research initiatives; moreover, through the SCRE control dashboard, it is possible to manually insert projects that are then processed by SCRE and indexed in the GoTriple index.

State the purpose of the data collection/generation

The TRIPLE core raw data is the starting point of the enrichment process and corresponds to the metadata of documents and projects collected from aggregators and providers. The collection happens directly through the OAI-PMH protocol, or indirectly through databases' dumps ingestion. In both cases, the collected metadata is mapped to the TRIPLE data model[2] used for indexing.

The raw metadata is then collected and enriched using controlled vocabularies in order to improve its quality and discoverability using different enrichment processes. The enriched metadata is subsequently indexed into the 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. It improves its quality, which is one of the objectives of the TRIPLE project. It is indexed in order to create the TRIPLE database. The TRIPLE data model is based on the schema.org vocabulary which enables interoperability between applications: recently it has been formally described in the TRIPLE Ontology[3]. In addition, search APIs 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 analysing their semantic proximity to different categories. These categories are taken from a multilingual thesaurus (described below).
- Normalisation using controlled vocabularies.
- Semantic annotation using the TRIPLE vocabulary
- Disambiguation tools to recognise duplicate documents and authors.

[2] See D2.2 Data harvesting best practices document for data providers. (Final)

[3] https://www.gotriple.eu/ontology/triple/index-en.html

In addition we use ElasticSearch indexing on the resulting data.

A multilingual disciplinary thesaurus for the SSH fields in ten languages - namely the *TRIPLE Vocabulary* - has been produced in the project by the WP2 team. The languages in question are Croatian, English, French, German, Greek, Italian, Polish, Portuguese, Spanish and Ukrainian. Additionally, the TRIPLE Vocabulary includes Dutch and Finnish labels which were retrieved from Library of Congress Subject Headings mappings. It is published as a Linked Open Data (LOD) dataset[4] in RDF (Resource Description Framework) using the Simple Knowledge Organisation System (SKOS), a W3C recommendation for the representation of Semantic Web controlled vocabularies.

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

The data from enrichment has been reused to continue improving search engine metadata especially for the purpose of machine learning. The indexed data has been used to create the TRIPLE database. Data is reused to connect with certain Innovative Services through a REST API.

Specify the origin of the data

TRIPLE raw data comes from national or international aggregators (e.g. OpenAIRE, DOAJ, Isidore) and from providers of any size in the European area. The raw data consists of reusable metadata following the DublinCore format or compliant formats (e.g. OpenAIRE metadata format).

Raw data combined with enrichment information counts as enriched data which is then indexed. For classification, the machine learning model has been trained with scholarly articles from journals referenced in the Directory of Open Access Journals (DOAJ)[5] and other relevant sources. Categorization, normalisation and semantic annotations are carried out using controlled vocabularies based on existing SSH catalogues[6].

State the expected size of the data (if known)

At the end of February 2023, The GoTriple platform contains more than 6.4 million documents, 5 million profiles and about 21 thousand projects. The current global size of the data is: 80Gb for the Elasticsearch data files for the front-end indexes and around 95Gb for the SCRE cache. The size of data is constantly increasing and will grow even after the end of the project. As of 28 February 2023, the total number of searchable data is over 11.5 million, including publications/datasets (documents), projects and profiles (people), as shown in the image below

[4] https://www.semantics.gr/authorities/vocabularies/SSH-LCSH/vocabulary-entries

[5] https://doaj.org/

[6] 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 catalogues) (annotation-disambiguation)

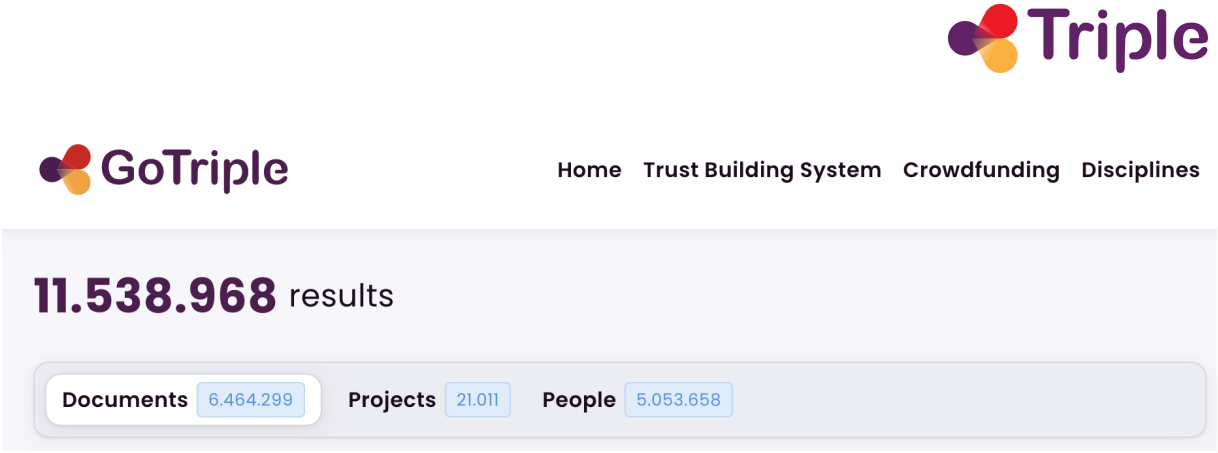

FIG.2 : GoTriple data at the end of February 2023

Outline the data utility: to whom it will 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. Through the acquisition of data, the platform is 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 have been 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 Machine Learning Data

> ***Dataset Description**: Machine Learning training data are in the form of metadata (title, keywords, abstracts) in eleven TRIPLE 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 documents in eleven TRIPLE languages (Croatian, English, French, German, Greek, Italian, Polish, Portuguese, Slovenian, Spanish and Ukrainian).

Explain the relation to the objectives of the project

Used to semantically enrich the Delivery Platform Raw Data with semantic classification against the MORESS categories.

Specify the types and formats of data generated/collected

Machine learning data is 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, or manually collected from other open sources by WP2 partners. These data are 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. Additional data were collected by the project partners who followed the data collection instructions annexed in the current version of the DMP (see Annex 1).

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

Outline the data utility: to whom will it be useful
The data is useful for users of the system as it is used to enrich metadata records and make resources more findable.
A deliverable in the form of a report on machine learning was submitted in December 2021[7] and provides precise details on the salient aspects of this data.

1.4 Innovative Services Data

Dataset Description: *The "Innovative Services'' are applications and tools that are not part of the core of the GoTriple platform. These applications and tools work on 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: a recommender system; a discovery and visualisation system; the Open annotation tool; the Trust Building System (TBS); Third-Party Applications, including a crowdfunding platform. Types of Innovative Services Data include:*[8]

- *User Interaction Data on the GoTriple platform (recommender system)*
- *Visual representations of GoTriple Platform data (Visualisation and Discovery System)*
- *Annotations applied on web resources, in particular web pages and .pdf documents, applied with the Pundit Open Annotation tool (https://thepund.it)*[9]
- *Data related to the Trust Building System (TBS).*

[7] See D2.3 Report on Machine Learning. .

[8] *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 are envisioned at present.*

[9] *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 an autonomous fashion by the Pundit service, which can be seen as "external" in relationship with the TRIPLE Core.*

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

State the purpose of the data collection/generation
User interaction data on the GoTriple Platform is tracked to form the basis for the personalised services of the project. Such services improve user experience, support in decision making and assistance in finding relevant items and peers.

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

Specify the types and formats of data generated/collected
User interactions on the GoTriple Platform are 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 is specified in accordance with the use case. It is for example the ID of the resource visited by the user together with its type, the action they performed on the site, etc).

Specify if existing data is being re-used (if any)
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 visualisations or any kind of behavioural or social modelling that may serve as a basis for such services. Additionally, end users will benefit from the improved User Experience.

1.4.2 Innovative Services Data: Discovery system and visualisation

State the purpose of the data collection/generation
To create map representations for the specific purpose of visual presentation of GoTriple search results.

Explain the relation to the objectives of the project

The service expands the User Interface of GoTriple with an advanced interactive search service, enriched with sophisticated visualisations. It is 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)
The Discovery System re-uses TRIPLE data indexed by BASE: the latter imports them periodically through the OAI-PMH protocol from the TRIPLE Core database.

Specify the origin of the data
Metadata from the TRIPLE core database that is imported by the BASE aggregator; 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.

1.4.3 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 and .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**).
- Association of a free-form text "tag" to an annotation (**oa:tagging**).
- 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 allows 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.

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 Elasticsearch indexes). Users are able to export their annotations in several open formats, including JSON and ODT. Moreover Pundit provides an open API that returns annotation data in a fully compliant format with the W3C Web annotation standard, serialised 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 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 reuses 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 around 107,000 annotations are stored in Pundit.

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.

1.4.4 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 are kept to their 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 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 (username, email address)
- Profile data (location, company, position, education, about)
- UGC: posts (AES encrypted)
- Chat messages (AES encrypted)

Formats: Proprietary
Workshop and Questionnaire data: same format as 'Platform co-design data'.

Specify if existing data is being re-used (if any)
Not applicable.

Specify the origin of the data
Not applicable.

State the expected size of the data (if known)
The size is approximately 350MB.

Outline the data utility: to whom will it be useful
To the users of the TBS; to the users of the GoTriple platform and its innovative services; to researchers; to the administrative team.

1.4.5 Third-Party Applications Data

Third party applications were selected and integrated in GoTriple in task 5.1. They are:

- a crowdfunding platform, operated by the company WeMakeIt
- the OPERAS Metrics service, operated by the company Ubiquity Press
- bookmarking services that GoTriple users can use to keep track of the publications found on the platform.

None of them is relevant to this DMP. The former two services are operated independently by companies external to the TRIPLE Consortium: data is therefore completely managed by them in their own infrastructures. GoTriple simply uses these services without any exchange of data. The Crowdfunding has been in fact integrated as a link to the WeMakeIt platform, while the visualisation of the OPERAS Metrics are produced by dynamically calling on the GoTriple front-end the Ubiquity Press APIs, sending as a key the DOI of a publication, if present.

Finally the bookmarking services integration consists of an export in BibTeX format of the publications data maintained in the GoTriple indexes. No specific data is generated and managed because of this service.

1.5 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 has been analysed to inform the building of research instruments (interviews, questionnaires) and has been included in the literature review, part of Deliverable D3.1[10].

[10] See D3.1 Iteration on the User Needs. (Final).

Specify the types and formats of data generated/collected
Bibliographic records have been collected and stored in the Zotero open-source bibliographic system. The Zotero account will be maintained to keep track of the main bibliographic records and may be reused for further projects related to the sustainability of the platform.

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 approximately 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 for stakeholders interested in user-focused research on digital practices within SSH.

1.6 Communication Data

> ***Dataset Description****: publications made on the TRIPLE project (powerpoints, abstracts, papers, workshops, webinars, posters, scientific articles, and video recordings). 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 to SSH researchers' communities to inform about the objectives and ongoing of the project. The Zenodo platform is used to disseminate the main reports and scientific publications of the TRIPLE project. Zenodo has been chosen because it is aligned with the Open Science and FAIR principles. It is also GDPR compliant, as stated in the Zenodo privacy policy

https://about.zenodo.org/privacy-policy/. By the way a step-by-step (Zenodo Upload Manual) is available in the repository of the project, enabling the WP8 partners to add resources in Zenodo, following the requirements such as for instance, the OPERAS community to be tagged, the keywords, the licences (**Access right**: "Open Access", **Licence**: "Creative Commons Attribution 4.0 International" (default setting)).

- 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, EGI conference 2020, CLARIN Annual Conference 2021, DARIAH Annual Conference: "Interfaces" 2021, Open Science Conference 2022 and IASSIST 2022 – Data by Design: Building a Sustainable Data Culture.
- A promotional video was shared on the TRIPLE Twitter account and broadcasted during the PUBMET event in September 2021.
- Video recordings and materials (powerpoints) of WP6 training sessions about Open Science are stored on the secured Nakala platform (maintained by Huma-Num) and deposited on Zenodo under the OPERAS Community. They are also disseminated through the DARIAH-Campus platform where they are hosted as external Learning Resources. Sessions are recorded via the Zoom application.
  The sessions are under the licence CC BY 4.0. Metadata disseminated are under the DARIAH-Campus Legal Notice[11] and Reuse Charter[12] to commit to the six core principles of Reciprocity, Interoperability, Citability, Openness, Stewardship and Trustworthiness and show how this commitment is reflected in the design and the daily operations of the platform.
- Video recordings and photos of demonstrations, presentations, tutorials and workshops about the TRIPLE project are broadcasted via the TRIPLE Youtube channel whose privacy policies are described in the section 5.2.

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. For video recordings, we use the commonly used format .mp4. 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 on Zenodo platform via a DOI or via a permanent url link.

[11] DARIAH-Campus Legal Notice: https://campus.dariah.eu/imprint
[12] DARIAH-Campus Reuse Charter: https://campus.dariah.eu/docs/reuse-charter

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 toolkits are accessible within the consortium to ensure harmonised communication data.

State the expected size of the data (if known)
The exact size is not known but remains 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 providers. This raw metadata describes documents, projects and researcher profiles. The enriched metadata is used for the purposes of data discovery and by innovative services.
For the specific case of co-design research data, much of the data generated isbe qualitative. As such, this data is not discoverable through automated means. Each project partner responsible for the user research data collection in WP3 stores the data on their own server (or chosen GDPR compliant service). Questionnaire data is stored with a copy of the questionnaire in WP3 leader's Secure Storage Drive so as to facilitate eventual data reuse.

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 and providers. As far as possible identifiers (DOI, Handle, ark and local identifier) for documents, research projects and researchers' profiles will be re-used when possible. PIDs are a mandatory requirement for all the resources harvested by GoTriple.

Outline naming conventions used
File naming conventions for TRIPLE data relate only to the internal processes of the TRIPLE system and are not 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 platform, an automatic update of the documents takes place every two months in the form of a 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 TRIPLE data model was built by using the schema.org[13] vocabulary and is essentially aligned with Dublin Core and the OpenAIRE metadata schemas. Other more specific mappings have been established (e.g. with the Europeana Data Model).

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

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 licences allowing reuse according to European Commission requirements. Hence, the platform has set standards by determining the rules for open research practices and workflows. Efforts were carried out to set up guidelines to harmonise Open Access and Open Science policies and practices among the various European organisations who participate in the platform in view of developing a shared vision which places Open Access and responsible research at the forefront. Direct and indirect identifiers are removed from public-use data files to minimise 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 APIs and OAI-PMH protocol for selected partners.
- TRIPLE Vocabulary: the SSH multilingual thesaurus of subjects that is created within task T2.4 that will be primarily used by the annotation service, has been published as LOD and via API under a CC licence. It is available by a LD endpoint, for downloading and via APIs.
- Integrated services data: the integrated services internal data are not 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 publications are available to the broader SSH research community under the CC-BY-4.0 open licence, in order to enable its reuse.

The TRIPLE data gathered (metadata on datasets, publications, profiles, etc.) is exploited in two ways:

- Internally in the form of data enrichment/refinement into innovative platform features (recommender services, visualisations, customised reports etc.)
- Externally via APIs to third party service providers for further developments (to this end GoTriple recommends to its providers the use of CC-0 licensed metadata).

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 current version also provides 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.

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 is included as well as the open source code of the relevant software. The selection of the repository for such documentation is on-going.

Specify where the data and associated metadata, documentation and code are deposited
TRIPLE data and code are managed by Net7 according to TRIPLE specifications. Each external service integrated in TRIPLE is 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
Regarding 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.
Metadata records produced by TRIPLE are published using the following standard vocabularies: schema.org, Dublin Core Metadata Element Set, OpenAIRE metadata schema, Europeana Data Model. Annotations data is serialised in JSON-LD according to the W3C Web Annotation standard specifications. The TRIPLE Vocabulary is published as RDF (both in RDF/XML and JSON-LD serialisation) and particularly according to the Simple Knowledge Organisation System (SKOS) ontology.

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

Specify how the data will be licenced to permit the widest reuse possible
Data Licence
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 licences allowing reuse, according to the Commission requirements. This also holds for the TRIPLE Vocabulary which has been published under an open licence (CC-BY-4.0).

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 GoTriple 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[14], shares the responsibility of Data management with the IT engineer in charge of the Data management of Net7 (responsible for the global architecture of the GoTriple platform).

Describe costs and potential value of long term preservation

The FAIRification process is part of the GoTriple platform development: data providers have to provide already FAIRified content, which is further enriched by GoTriple 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. Two actions have been taken for that purpose: a “Handbook for GoTriple Content Providers” has been established, providing useful information about GoTriple’s requirements and guidance about FAIR principles implementation; a harvesting management system (HMS) has been set up for the providers to assess both the data and the harvesting quality of the contents’ ingestion.

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

3. DATA EXPLOITATION

Usage metrics are collected from three different dashboards, the database, Mixpanel and Matomo.) These anonymized usage metrics provide useful information on the platform's usability, dynamics, and potential enhancements; the usage metrics are 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 are quantitative data similar to the quantitative data generated for WP3. The usage metrics have an important role in ensuring the sustainability and soundness of GoTriple's business model.

4. DATA SECURITY

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

Relevant data from TRIPLE is deposited within the central digital project repository Sharedocs run by CNRS (Huma-Num) to ensure that the research community has long-term access to the data. In order to do so, the capabilities of TRIPLE and its trained archival staff were leveraged. CNRS (Huma-Num) has a strong expertise in preservation and storage. To avoid the loss of data, CNRS (Huma-Num) makes use of appropriate formats in order to ensure data interoperability, facilitate the archiving process and make the storage of data independent of the device used to disseminate the data. CNRS (Huma-Num) provides a long-term preservation service based on the CINES[15] facility (archiving), which is intended for data with a valuable heritage or scientific value.

Throughout the life of the project, TRIPLE ensures 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 organisation 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.

The secure storage provisions for the other sources of data are listed below:

- Co-design research data: much of the data is stored on the WP3 leader Secure Research drive (this includes data from tasks 3.1, 3.2 and most of 3.6). This ensures 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 is stored in an offline secure data storage.
- The data collected by the partner OKMaps during their workshop has been processed and stored by OKMaps and TRIPLE.
- The data collected by EKT for Task 3.5 is 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, are used. All data is anonymized after completion of the workshop.
- The TRIPLE vocabulary is stored in the Semantics.gr platform, a platform developed by EKT for managing and publishing vocabularies, thesauri and authority files. The platform along with the hosting data (vocabularies) are safely stored in EKT Data Centre. EKT has a strong expertise in preservation and storage.
- The video recordings of the Open Science training sessions are stored and publicly accessible in Sharedocs, an application hosted by Huma-Num that allows researchers to share, publish and promote all types of documented digital data (text files, sounds, images, videos, 3D objects, etc.). This website, which contains personal information concerning the staff of the CNRS and its partners, has been declared to the CNIL (notice in

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

the process of being filed). In accordance with the law n° 78-17 of January 6, 1978, relating to Data processing, the files and Freedoms (articles 38, 39, 40), you have a right of access, of correction and suppression of the data concerning you, on line on this site. Videos are licensed under CC-BY-4.0.

Some of the Co-design research data are 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., Abertay University uses a secured Drive for this purpose, which ensures compliance with GDPR and protection of the data. In the case of 2., the service is GDPR compliant. In the case of 3., we use the TRIPLE project secured 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. It Includes references and related technical aspects.

Informed consent: For TRIPLE project, 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 project ensures that personal data processing and management respect the General Data Protection Regulation (GDPR) provisions, by adopting a privacy by design approach. TRIPLE website's privacy policy is described in a specific document that is publicly accessible on the project website and annexed in this version of the DMP (See Annex 2). Personal data is collected for the compilation of individual profiles. In this case, data such as first name, surname, encrypted identifiers and IP address are used to enable the social network functionality which will be part of the GoTriple services. Third-party personal data processing (e.g. interoperable identifiers like ORCID) will depend on their privacy policy. Users receive clear information when using the service and are informed of their rights. Other personal data is automatically collected for the purposes of metrics, especially through the use of cookies. This enables measurements of site traffic and usage. A privacy policy document gives more details about the duration of personal data storage (See Annex 2), but storage for metrics purposes will not exceed 12 months. The responsible for processing in TRIPLE project will be the Project Coordination Team (PCT).

5.1 GoTriple communities mailing list

A GoTriplecommunity mailing list (gotriple-communities@listes.huma-num.fr) was set up and managed by Huma-Num. The mailing list enters the framework of a commercial prospecting BtoC from the CNIL[16], the French National Data Regulation Centre.
Obligations are therefore

- Obtain clear consent:
  - No pre-checked boxes and double opt-in are recommended because it is necessary to keep the proof of this consent (and a double opt-in is a reinforced proof).
- For consent to be informed and clear, the following information must be provided:

[16] CNIL: https://www.cnil.fr/en/home

- Purpose of the collection: purpose of the mailing list and therefore of the data collection
- Clear information on how to unsubscribe: include a functional link
- Clear information on the issuer: name of the organisation / postal address / e-mail / name / DPO

- Provide updated information notices: TOS / data management policy / legal notice (including user rights: access / portability / rectification / deletion).
- On the side of the issuing organisation it is necessary to prove consent at any time (so store these "yeses" somewhere)
- Following information is systematically included in each email: (1) information about the identity of the advertiser and (2) a link to unsubscribe.

This mailing list is used exclusively by WP3 to test the first release of the platform. As of 9 March 2022, this mailing list includes 62 volunteers to test the platform and participate in face to face interviews. The volunteers consented in a formal way via a google registration form and gave their written consent for the processing of their personal data for the purposes of receiving information on their email address regarding the project and accordingly to the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

The information stored are the email addresses, no other data are required and only 2 persons have access to the list, one person from Huma-Num as the owner of the mailing list and one person from WP3 to contact volunteers for face to face interviews.

5.2 TRIPLE Youtube channel[17]

This channel was created in June 2021 to broadcast the records of TRIPLE workshops and presentations of the project and of the platform. Currently, 20 videos are disseminated through the Youtube channel which contains:

- The WP6 Open Science training series
- The presentations of the first TRIPLE conference
- Video demonstrations of the GoTriple discovery platform
- TRIPLE Video tutorials
- TRIPLE Webinars
- OPERAS Crowdfunding workshops

Videos are under the Google privacy policies[18].

[17] https://www.youtube.com/channel/UCjwEHdltYYhocCC9o6RljuQ

[18] https://policies.google.com/privacy?hl=en-US andhttps://www.youtube.com/static?template=terms

The internal TRIPLE privacy policy is accessible via the TRIPLE website and the link is provided in the "About us" tab of the Youtube channel: https://project.gotriple.eu/privacy-policy, the policy was updated in August 2021. The legal notice is accessible publicly on the project website to the following link: https://project.gotriple.eu/contact.

The videos are licensed under CC-BY-4.0.

5.3 TRIPLE website

5.3.1. Privacy policy

The TRIPLE privacy policy is accessible to the following link : https://project.gotriple.eu/privacy-policy/, and annexed to version 2 of the DMP (See Annex 2).

5.3.2 Terms of use

The use of the TRIPLE website constitutes an agreement with the following terms and conditions:

(a) The TRIPLE consortium maintains the website as a courtesy to those who may choose to access the Site ("Users"). The information presented is for informative purposes only. Unless otherwise indicated, all materials created by the OPERAS network are licensed under a Creative Commons Attribution 4.0 International Licence.

(b) Huma-Num and Max Weber Foundation – German Humanities Institutes Abroad (MWS) – as part of the TRIPLE consortium – administer this Site. All material on this site appears subject to the terms and conditions described on the website. Unless noted otherwise all content is licenced under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

5.3.3 Data processing agreement

The processing is in accordance with Article 28 General Data Protection Regulation (GDPR) (as of: June 2021). The subject-matter of the Agreement regarding the processing of data is the execution of the following services or tasks by the Processor website hosting, TRIPLE Community Mailing List and Matomo analytics tool. The Agreement is between the Data controller (MWS) and the Data processor (Huma-Num). It is currently being updated to guarantee the long term sustainability of the agreement and thereby the project website.

Annex 1 Data Collection instructions

Types of data

journal publications, conference publications, books, blogs, theses, datasets, projects
Minimum requirements: **title** and **abstract**
Very useful: **keywords**
A valid data entry must contain title and abstract. Every effort should be made to include keywords as well but the entry is valid even without them. A full text of the paper is not required.

Task

Collect enough academic papers so that their number exceeds 130 per category (together with the already collected DOAJ papers.)

Paper selection

Your spreadsheet for collecting metadata contains a list of MORESS categories and the number of papers downloaded from DOAJ to show how many are missing to make a total of **130 per category**.
Use the comments in DOAJ evaluation and select papers so that they balance any bias in the DOAJ papers and add the missing topics that also belong to the category.
It is safest to start with categories where there is nothing from DOAJ or just a couple of papers.

Collection folder

Both spreadsheets for metadata and folders with the name of the language are in the TRIPLE Google Drive folder called **Data collection**.
Access to the folder: The folder is located in the standard TRIPLE folder. If you have access to the TRIPLE folder, you have access to the collection folder too. If you need access, contact other TRIPLE people to grant you access.

How

Register metadata

Use your local expertise to find suitable sources, identify suitable papers. Register the metadata (and optionally, the file name) in the spreadsheet. Rename the file if needed to make the file name unique.

Questions

Contact: Ondřej ondrej.matuska@sketchengine.eu

Bulk downloads

If you locate a large repository of suitable papers, Lexical Computing can download the papers for you in bulk. This only makes sense if there are hundreds of papers to download.
Otherwise the manual procedure is more effective.
The procedure will be agreed on a case by case basis. Please use the above contacts to discuss the options.

Annex 2 TRIPLE Website Privacy Policy

Preamble

With the following privacy policy we would like to inform you which types of your personal data (hereinafter also abbreviated as “data”) we process for which purposes and in which scope. The privacy statement applies to all processing of personal data carried out by us, both in the context of providing our services and in particular on our websites, in mobile applications and within external online presences, such as our social media profiles (hereinafter collectively referred to as “online services”).

The terms used are not gender-specific.

Last Update: 2. August 2021

Table of contents

Controller

This website is administered by the Max Weber Foundation – German Humanities Institutes Abroad (MWS) – on behalf of the TRIPLE consortium.

Max Weber Foundation – German Humanities Institutes Abroad

Central Office

Rheinallee 6

53173 Bonn

Germany

Tel. +49-(0)228-37786-0

Email: triple[at]operas-eu.org

Website: https://www.maxweberstiftung.de

Authorised Representatives: Sona Arasteh (TRIPLE Communication Officer, work package 8 leader and chief editor of the TRIPLE project website).

E-mail address: arasteh@maxweberstiftung.de.

Legal Notice: https://project.gotriple.eu/contact/

Contact Information of the Data Protection Officer

Reinhard Hiss - datenschutz@maxweberstiftung.de

Overview of Processing Operations

Below is a summary of the types of data processed, the purposes for which they are processed and the concerned data subjects.

Categories of Processed Data

- Inventory data (e.g. names, addresses).
- Content data (e.g. text input, photographs, videos).
- Contact data (e.g. e-mail, telephone numbers).
- Meta/communication data (e.g. device information, IP addresses).
- Usage data (e.g. websites visited, interest in content, access times).
- Contract data (e.g. contract object, duration, customer category).

- Payment Data (e.g. bank details, invoices, payment history).

Categories of Data Subjects

- Business and contractual partners.
- Communication partner (Recipients of emails, letters, etc.).
- Members.
- Users (e.g. website visitors, users of online services).

Purposes of Processing

- Provision of our online services and usability.
- Direct marketing (e.g. by e-mail or postal).
- Feedback (e.g. collecting feedback via online form).
- Marketing.
- Contact requests and communication.
- Profiles with user-related information (Creating user profiles).
- Web Analytics (e.g. access statistics, recognition of returning visitors).
- Security measures.
- Provision of contractual services and customer support.
- Managing and responding to inquiries.

Legal Bases for the Processing

In the following section, you will find an overview of the legal basis of the GDPR on which we base the processing of personal data. Please note that in addition to the provisions of the GDPR, national data protection provisions of your or our country of residence or domicile may apply. If, in addition, more specific legal bases are applicable in individual cases, we will inform you of these in the data protection declaration.

- Consent (Article 6 (1) (a) GDPR) – The data subject has given consent to the processing of his or her personal data for one or more specific purposes.
- Performance of a contract and prior requests (Article 6 (1) (b) GDPR) – Performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract.
- Legitimate Interests (Article 6 (1) (f) GDPR) – Processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where

such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data.

National data protection regulations in Germany: In addition to the data protection regulations of the General Data Protection Regulation, national regulations apply to data protection in Germany. This includes in particular the Law on Protection against Misuse of Personal Data in Data Processing (Federal Data Protection Act – BDSG). In particular, the BDSG contains special provisions on the right to access, the right to erase, the right to object, the processing of special categories of personal data, processing for other purposes and transmission as well as automated individual decision-making, including profiling. Furthermore, it regulates data processing for the purposes of the employment relationship (§ 26 BDSG), in particular with regard to the establishment, execution or termination of employment relationships as well as the consent of employees. Furthermore, data protection laws of the individual federal states may apply.

Security Precautions

We take appropriate technical and organisational measures in accordance with the legal requirements, taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, in order to ensure a level of security appropriate to the risk.

The measures include, in particular, safeguarding the confidentiality, integrity and availability of data by controlling physical and electronic access to the data as well as access to, input, transmission, securing and separation of the data. In addition, we have established procedures to ensure that data subjects' rights are respected, that data is erased, and that we are prepared to respond to data threats rapidly. Furthermore, we take the protection of personal data into account as early as the development or selection of hardware, software and service providers, in accordance with the principle of privacy by design and privacy by default.

Masking of the IP address: If IP addresses are processed by us or by the service providers and technologies used and the processing of a complete IP address is not necessary, the IP address is shortened (also referred to as "IP masking"). In this process, the last two digits or the last part of the IP address after a full stop are removed or replaced by wildcards. The masking of the IP address is intended to prevent the identification of a person by means of their IP address or to make such identification significantly more difficult.

SSL encryption (https): In order to protect your data transmitted via our online services in the best possible way, we use SSL encryption. You can recognize such encrypted connections by the prefix https:// in the address bar of your browser.

Erasure of Data

The data processed by us will be erased in accordance with the statutory provisions as soon as their processing is revoked or other permissions no longer apply (e.g. if the purpose of processing this data no longer applies or they are not required for the purpose).

If the data is not deleted because they are required for other and legally permissible purposes, their processing is limited to these purposes. This means that the data will be restricted and not processed for other purposes. This applies, for example, to data that must be stored for commercial or tax reasons or for which storage is necessary to assert, exercise or defend legal claims or to protect the rights of another natural or legal person.

In the context of our information on data processing, we may provide users with further information on the deletion and retention of data that is specific to the respective processing operation.

Use of Cookies

Cookies are text files that contain data from visited websites or domains and are stored by a browser on the user's computer. A cookie is primarily used to store information about a user during or after his visit within an online service. The information stored can include, for example, the language settings on a website, the login status, a shopping basket or the location where a video was viewed. The term "cookies" also includes other technologies that fulfil the same functions as cookies (e.g. if user information is stored using pseudonymous online identifiers, also referred to as "user IDs").

The following types and functions of cookies are distinguished:

- Temporary cookies (also: session cookies): Temporary cookies are deleted at the latest after a user has left an online service and closed his browser.
- Permanent cookies: Permanent cookies remain stored even after closing the browser. For example, the login status can be saved or preferred content can be displayed directly when the user visits a website again. The interests of users who are used for range measurement or marketing purposes can also be stored in such a cookie.
- First-Party-Cookies: First-Party-Cookies are set by ourselves.
- Third party cookies: Third party cookies are mainly used by advertisers (so-called third parties) to process user information.
- Necessary (also: essential) cookies: Cookies can be necessary for the operation of a website (e.g. to save logins or other user inputs or for security reasons).
- Statistics, marketing and personalisation cookies: Cookies are also generally used to measure a website's reach and when a user's interests or behaviour (e.g. viewing certain content, using functions, etc.) are stored on individual websites in a user

profile. Such profiles are used, for example, to display content to users that corresponds to their potential interests. This procedure is also referred to as “tracking”, i.e. tracking the potential interests of users. If we use cookies or “tracking” technologies, we will inform you separately in our privacy policy or in the context of obtaining consent.

Information on legal basis: The legal basis on which we process your personal data with the help of cookies depends on whether we ask you for your consent. If this applies and you consent to the use of cookies, the legal basis for processing your data is your declared consent. Otherwise, the data processed with the help of cookies will be processed on the basis of our legitimate interests (e.g. in a business operation of our online service and its improvement) or, if the use of cookies is necessary to fulfil our contractual obligations.

Retention period: Unless we provide you with explicit information on the retention period of permanent cookies (e.g. within the scope of a so-called cookie opt-in), please assume that the retention period can be as long as two years.

General information on Withdrawal of consent and objection (Opt-Out): Respective of whether processing is based on consent or legal permission, you have the option at any time to object to the processing of your data using cookie technologies or to revoke consent (collectively referred to as “opt-out”). You can initially explain your objection using the settings of your browser, e.g. by deactivating the use of cookies (which may also restrict the functionality of our online services). An objection to the use of cookies for online marketing purposes can be raised for a large number of services, especially in the case of tracking, via the websites https://www.aboutads.info/choices/ and https://www.youronlinechoices.com. In addition, you can receive further information on objections in the context of the information on the used service providers and cookies.

Processing Cookie Data on the Basis of Consent: We use a cookie management solution in which users’ consent to the use of cookies, or the procedures and providers mentioned in the cookie management solution, can be obtained, managed and revoked by the users. The declaration of consent is stored so that it does not have to be retrieved again and the consent can be proven in accordance with the legal obligation. Storage can take place server-sided and/or in a cookie (so-called opt-out cookie or with the aid of comparable technologies) in order to be able to assign the consent to a user or and/or his/her device.Subject to individual details of the providers of cookie management services, the following information applies: The duration of the storage of the consent can be up to two years. In this case, a pseudonymous user identifier is formed and stored with the date/time of consent, information on the scope of the consent (e.g. which categories of cookies and/or service providers) as well as the browser, system and used end device.

- Processed data types: Usage data (e.g. websites visited, interest in content, access times), Meta/communication data (e.g. device information, IP addresses).
- Data subjects: Users (e.g. website visitors, users of online services).

- Legal Basis: Consent (Article 6 (1) (a) GDPR), Legitimate Interests (Article 6 (1) (f) GDPR).

Performing tasks in accordance with statutes or rules of procedure

We process the data of our members, supporters, prospects, business partners or other persons (collectively, ” data subjects “) when we have a membership or other business relationship with them and perform our functions and are recipients of benefits and benefits. Otherwise, we process the data of data subjects on the basis of our legitimate interests, e.g. when it concerns administrative tasks or public relations.

The data processed, the type, scope and purpose and the necessity of their processing, are determined by the underlying membership or contractual relationship, from which the necessity of any data information arises (otherwise we refer to necessary data).

We delete data that is no longer required for the performance of our statutory and business purposes. This is determined according to the respective tasks and contractual relationships. We retain the data for as long as it may be relevant for the purpose of conducting business and with regard to any warranty or liability obligations on the basis of our legitimate interest in their regulation. The necessity of storing the data is checked regularly; otherwise the statutory storage obligations apply.

- Processed data types: Inventory data (e.g. names, addresses), Payment Data (e.g. bank details, invoices, payment history), Contact data (e.g. e-mail, telephone numbers), Contract data (e.g. contract object, duration, customer category).
- Data subjects: Users (e.g. website visitors, users of online services), Members, Business and contractual partners.
- Purposes of Processing: Provision of contractual services and customer support, Contact requests and communication, Managing and responding to inquiries.
- Legal Basis: Performance of a contract and prior requests (Article 6 (1) (b) GDPR), Legitimate Interests (Article 6 (1) (f) GDPR).

Provision of online services and web hosting

In order to provide our online services securely and efficiently, we use the services of one or more web hosting providers from whose servers (or servers they manage) the online services can be accessed. For these purposes, we may use infrastructure and platform services, computing capacity, storage space and database services, as well as security and technical maintenance services.

The data processed within the framework of the provision of the hosting services may include all information relating to the users of our online services that is collected in the course of use and communication. This regularly includes the IP address, which is necessary to be able to deliver the contents of online services to browsers, and all entries made within our online services or from websites.

Collection of Access Data and Log Files: We, ourselves or our web hosting provider, collect data on the basis of each access to the server (so-called server log files). Server log files may include the address and name of the web pages and files accessed, the date and time of access, data volumes transferred, notification of successful access, browser type and version, the user's operating system, referrer URL (the previously visited page) and, as a general rule, IP addresses and the requesting provider.

The server log files can be used for security purposes, e.g. to avoid overloading the servers (especially in the case of abusive attacks, so-called DDoS attacks) and to ensure the stability and optimal load balancing of the servers.

- Processed data types: Content data (e.g. text input, photographs, videos), Usage data (e.g. websites visited, interest in content, access times), Meta/communication data (e.g. device information, IP addresses).
- Data subjects: Users (e.g. website visitors, users of online services).
- Purposes of Processing: Provision of our online services and usability.
- Legal Basis: Legitimate Interests (Article 6 (1) (f) GDPR).

Blogs and publication media

We use blogs or comparable means of online communication and publication (hereinafter "publication medium"). Readers' data will only be processed for the purposes of the publication medium to the extent necessary for its presentation and communication between authors and readers or for security reasons. For the rest, we refer to the information on the processing of visitors to our publication medium within the scope of this privacy policy.

By default, the comment function is disabled on the blog (called "News" on the website).

Akismet Anti-Spam Checking: We use the "Akismet" service on the basis of our legitimate interests. With the help of Akismet, comments from real people are distinguished from spam comments. All comments are sent to a server in the USA, where they are analysed and stored for four days for comparison purposes. If a comment has been classified as spam, the data will be stored beyond that time. This information includes the name entered, the e-mail address, the IP address, the comment content, the referrer, information about the browser used, the computer system and the time of the entry.

Users are welcome to use pseudonyms, or to refrain from entering their name or email address. You can completely prevent the transmission of data by not using our comment

system. That is a pity, but unfortunately we do not see any alternatives that work just as effectively.

- Processed data types: Inventory data (e.g. names, addresses), Contact data (e.g. e-mail, telephone numbers), Content data (e.g. text input, photographs, videos), Usage data (e.g. websites visited, interest in content, access times), Meta/communication data (e.g. device information, IP addresses).
- Data subjects: Users (e.g. website visitors, users of online services).
- Purposes of Processing: Provision of contractual services and customer support, Feedback (e.g. collecting feedback via online form), Security measures.
- Legal Basis: Performance of a contract and prior requests (Article 6 (1) (b) GDPR), Legitimate Interests (Article 6 (1) (f) GDPR).

Services and service providers being used:

- Akismet Anti-Spam Checking: Akismet Anti-Spam Checking; Service provider: Automattic Inc., 60 29th Street #343, San Francisco, CA 94110, USA; Website: https://automattic.com; Privacy Policy: https://automattic.com/privacy.
- Wordfence: Firewall and security as well as error detection functions ; Service provider: Defiant, Inc., 800 5th Ave Ste 4100, Seattle, WA 98104, USA; Website: https://www.wordfence.com; Privacy Policy: https://www.wordfence.com/privacy-policy/; Standard Contractual Clauses (Safeguarding the level of data protection when processing data in third countries): https://www.wordfence.com/gdpr/dpa.pdf.

Contact and enquiry management

When contacting us (e.g. by contact form, e-mail, telephone or via social media), the data of the inquiring persons are processed insofar as this is necessary to answer the contact enquiries and any requested activities.

The response to contact enquiries within the framework of contractual or pre-contractual relationships is made in order to fulfil our contractual obligations or to respond to (pre)contractual enquiries and otherwise on the basis of the legitimate interests in responding to the enquiries.

- Processed data types: Inventory data (e.g. names, addresses), Contact data (e.g. e-mail, telephone numbers), Content data (e.g. text input, photographs, videos).
- Data subjects: Communication partner (Recipients of emails, letters, etc.).
- Purposes of Processing: Contact requests and communication.

- Legal Basis: Performance of a contract and prior requests (Article 6 (1) (b) GDPR), Legitimate Interests (Article 6 (1) (f) GDPR).

Web Analysis, monitoring and optimization

Web analysis is used to evaluate the visitor traffic on our website and may include the behaviour, interests or demographic information of users, such as age or gender, as pseudonymous values. With the help of web analysis we can e.g. recognize, at which time our online services or their functions or contents are most frequently used or requested repeatedly, as well as which areas require optimization.

In addition to web analysis, we can also use test procedures, e.g. to test and optimise different versions of our online services or their components.

For these purposes, so-called user profiles can be created and stored in a file (so-called “cookie”) or similar procedures in which the relevant user information for the aforementioned analyses is stored. This information may include, for example, content viewed, web pages visited and elements and technical data used there, such as the browser used, computer system used and information on times of use. If users have consented to the collection of their location data, these may also be processed, depending on the provider.

The IP addresses of the users are also stored. However, we use any existing IP masking procedure (i.e. pseudonymisation by shortening the IP address) to protect the user. In general, within the framework of web analysis, A/B testing and optimisation, no user data (such as e-mail addresses or names) is stored, but pseudonyms. This means that we, as well as the providers of the software used, do not know the actual identity of the users, but only the information stored in their profiles for the purposes of the respective processes.

Information on legal basis: If we ask the users for their consent to the use of third party providers, the legal basis of the processing is consent. Furthermore, the processing can be a component of our (pre)contractual services, provided that the use of the third party was agreed within this context. Otherwise, user data will be processed on the basis of our legitimate interests (i.e. interest in efficient, economic and recipient friendly services). In this context, we would also like to refer you to the information on the use of cookies in this privacy policy.

- Processed data types: Usage data (e.g. websites visited, interest in content, access times), Meta/communication data (e.g. device information, IP addresses).
- Data subjects: Users (e.g. website visitors, users of online services).
- Purposes of Processing: Web Analytics (e.g. access statistics, recognition of returning visitors), Profiles with user-related information (Creating user profiles).
- Security measures: IP Masking (Pseudonymization of the IP address).
- Legal Basis: Consent (Article 6 (1) (a) GDPR), Legitimate Interests (Article 6 (1) (f) GDPR).

Services and service providers being used:

- Matomo: The information generated by the cookie about your use of this website will only be stored on our server and not disclosed to third parties; Service provider: Web analytics/ reach measurement in self-hosting; Website: https://matomo.org/; Retention period: The cookies have a maximum storage period of 13 months.

Profiles in social networks (social media)

We maintain online presences within social networks and process user data in this context in order to communicate with the users active there or to offer information about us.

We would like to point out that user data may be processed outside the European Union. This may entail risks for users, e.g. by making it more difficult to enforce users' rights.

In addition, user data is usually processed within social networks for market research and advertising purposes. For example, user profiles can be created on the basis of user behaviour and the associated interests of users. The user profiles can then be used, for example, to place advertisements within and outside the networks which are presumed to correspond to the interests of the users. For these purposes, cookies are usually stored on the user's computer, in which the user's usage behaviour and interests are stored. Furthermore, data can be stored in the user profiles independently of the devices used by the users (especially if the users are members of the respective networks or will become members later on).

For a detailed description of the respective processing operations and the opt-out options, please refer to the respective data protection declarations and information provided by the providers of the respective networks.

Also in the case of requests for information and the exercise of rights of data subjects, we point out that these can be most effectively pursued with the providers. Only the providers have access to the data of the users and can directly take appropriate measures and provide information. If you still need help, please do not hesitate to contact us.

- Processed data types: Contact data (e.g. e-mail, telephone numbers), Content data (e.g. text input, photographs, videos), Usage data (e.g. websites visited, interest in content, access times), Meta/communication data (e.g. device information, IP addresses).
- Data subjects: Users (e.g. website visitors, users of online services).
- Purposes of Processing: Contact requests and communication, Feedback (e.g. collecting feedback via online form), Marketing.
- Legal Basis: Legitimate Interests (Article 6 (1) (f) GDPR).

Services and service providers being used:

- Twitter: Social network; Service provider: Twitter International Company, One Cumberland Place, Fenian Street, Dublin 2 D02 AX07, Ireland, parent company: Twitter Inc., 1355 Market Street, Suite 900, San Francisco, CA 94103, USA; Privacy Policy: https://twitter.com/de/privacy, (Settings) https://twitter.com/personalization.

Plugins and embedded functions and content

Within our online services, we integrate functional and content elements that are obtained from the servers of their respective providers (hereinafter referred to as "third-party providers"). These may, for example, be graphics, videos or city maps (hereinafter uniformly referred to as "Content").

The integration always presupposes that the third-party providers of this content process the IP address of the user, since they could not send the content to their browser without the IP address. The IP address is therefore required for the presentation of these contents or functions. We strive to use only those contents, whose respective offerers use the IP address only for the distribution of the contents. Third parties may also use so-called pixel tags (invisible graphics, also known as "web beacons") for statistical or marketing purposes. The "pixel tags" can be used to evaluate information such as visitor traffic on the pages of this website. The pseudonymous information may also be stored in cookies on the user's device and may include technical information about the browser and operating system, referring websites, visit times and other information about the use of our website, as well as may be linked to such information from other sources.

Information on legal basis: If we ask users for their consent (e.g. in the context of a so-called "cookie banner consent"), the legal basis for processing is this consent. Otherwise, user data will be processed on the basis of our legitimate interests (i.e. interest in the analysis, optimisation and economic operation of our online services). We refer you to the note on the use of cookies in this privacy policy.

- Processed data types: Usage data (e.g. websites visited, interest in content, access times), Meta/communication data (e.g. device information, IP addresses), Inventory data (e.g. names, addresses), Contact data (e.g. e-mail, telephone numbers), Content data (e.g. text input, photographs, videos).
- Data subjects: Users (e.g. website visitors, users of online services).
- Purposes of Processing: Provision of our online services and usability, Marketing, Profiles with user-related information (Creating user profiles).
- Legal Basis: Consent (Article 6 (1) (a) GDPR), Legitimate Interests (Article 6 (1) (f) GDPR), Performance of a contract and prior requests (Article 6 (1) (b) GDPR).

Services and service providers being used:

- Twitter plugins and contents: Twitter plugins and buttons – This can include content such as images, videos or text and buttons with which users can share content from this online service within Twitter. Service provider: Twitter International Company, One Cumberland Place, Fenian Street, Dublin 2 D02 AX07, Ireland, parent company: Twitter Inc., 1355 Market Street, Suite 900, San Francisco, CA 94103, USA; Website: https://twitter.com; Privacy Policy: https://twitter.com/en/privacy.
- YouTube videos: Video contents; Service provider: Google Ireland Limited, Gordon House, Barrow Street, Dublin 4, Ireland, , parent company: Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA; Website: https://www.youtube.com; Privacy Policy: https://policies.google.com/privacy; Opt-Out: Opt-Out-Plugin: https://tools.google.com/dlpage/gaoptout?hl=en, Settings for the Display of Advertisements: https://adssettings.google.com/authenticated.

Changes and updates to the privacy policy

We kindly ask you to inform yourself regularly about the contents of our data protection declaration. We will adjust the privacy policy as changes in our data processing practices make this necessary. We will inform you as soon as the changes require your cooperation (e.g. consent) or other individual notification.

If we provide addresses and contact information of companies and organisations in this privacy policy, we ask you to note that addresses may change over time and to verify the information before contacting us.

Rights of data subjects

As data subject, you are entitled to various rights under the GDPR, which arise in particular from Articles 15 to 21 of the GDPR:

- Right to Object: You have the right, on grounds arising from your particular situation, to object at any time to the processing of your personal data which is based on letter (e) or (f) of Article 6(1) GDPR, including profiling based on those provisions. Where personal data are processed for direct marketing purposes, you have the right to object at any time to the processing of the personal data concerning you for the purpose of such marketing, which includes profiling to the extent that it is related to such direct marketing.
- Right of withdrawal for consents: You have the right to revoke consents at any time.
- Right of access: You have the right to request confirmation as to whether the data in question will be processed and to be informed of this data and to receive further information and a copy of the data in accordance with the provisions of the law.

- Right to rectification: You have the right, in accordance with the law, to request the completion of the data concerning you or the rectification of the incorrect data concerning you.
- Right to Erasure and Right to Restriction of Processing: In accordance with the statutory provisions, you have the right to demand that the relevant data be erased immediately or, alternatively, to demand that the processing of the data be restricted in accordance with the statutory provisions.
- Right to data portability: You have the right to receive data concerning you which you have provided to us in a structured, common and machine-readable format in accordance with the legal requirements, or to request its transmission to another controller.
- Complaint to the supervisory authority: In accordance with the law and without prejudice to any other administrative or judicial remedy, you also have the right to lodge a complaint with a data protection supervisory authority, in particular a supervisory authority in the Member State where you habitually reside, the supervisory authority of your place of work or the place of the alleged infringement, if you consider that the processing of personal data concerning you infringes the GDPR.

Terminology and Definitions

This section provides an overview of the terms used in this privacy policy. Many of the terms are drawn from the law and defined mainly in Article 4 GDPR. The legal definitions are binding. The following explanations, on the other hand, are intended above all for the purpose of comprehension. The terms are sorted alphabetically.

- Controller: “Controller” means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.
- IP Masking: IP masking is a method by which the last octet, i.e. the last two numbers of an IP address, are deleted so that the IP address alone can no longer be used to uniquely identify a person. IP masking is therefore a means of pseudonymisation processing methods, particularly in online marketing.
- Personal Data: “personal data” means any information relating to an identified or identifiable natural person (“data subject”); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.
- Processing: The term “processing” covers a wide range and practically every handling of data, be it collection, evaluation, storage, transmission or erasure.

- Profiles with user-related information: The processing of “profiles with user-related information”, or “profiles” for short, includes any kind of automated processing of personal data that consists of using these personal data to analyse, evaluate or predict certain personal aspects relating to a natural person (depending on the type of profiling, this may include different information concerning demographics, behaviour and interests, such as interaction with websites and their content, etc.) (e.g. interests in certain content or products, click behaviour on a website or location). Cookies and web beacons are often used for profiling purposes.
- Web Analytics: Web Analytics serves the evaluation of visitor traffic of online services and can determine their behaviour or interests in certain information, such as content of websites. With the help of web analytics, website owners, for example, can recognize at what time visitors visit their website and what content they are interested in. This allows them, for example, to optimise the content of the website to better meet the needs of their visitors. For purposes of web analytics, pseudonymous cookies and web beacons are frequently used in order to recognise returning visitors and thus obtain more precise analyses of the use of an online service.

Annex 3 Protocol for designing and delivering online training materials - Sustainability document

Introduction

This document summarises Open Sciences and data management practices implemented within the subtask (T6.3.) of the TRIPLE project, dedicated to delivering online training events to the project participants and beyond, with special focus on scholars and research support personnel working in the field of Social Sciences and Humanities.

The resources and activities summarised in this document have been carried out in Work Package 6 “Open Science and the EOSC integration”.

Overview of resources

To consolidate good practice with the creation of FAIR-by-design online training events and materials, the task force has made not only the training materials available (slides, video recordings) but also supplementary materials that guided the process. These are listed in the first table and have been deposited in Zenodo in the form of a TRIPLE Training Toolkit (https://doi.org/10.5281/zenodo.7309919). The documents have been produced to support trainers and organisers of online training events and include in particular: 1) guidelines for the organisation of online training events, including a step-by-step checklist which can be reused by anyone wishing to organise an online training event and 2) survey instruments that trainers and organisers can reuse should they wish to assess the training needs and training impact in their community.

1. Auxiliary materials to give context and documentation to the training materials, guide the preparation and production of similar training events and share good practices of producing and disseminating FAIR-by-design training resources

	Name/Title of document	Format	Access type	Licence	Location
1	README	.docx	OA	CC-BY-4.0	Zenodo
2	TRIPLE_Training_Toolkit_Workflow1	.png	OA	CC-BY-4.0	Zenodo
3	TRIPLE_Training_Toolkit_Workflow2	.png	OA	CC-BY-4.0	Zenodo
4	Guidelines_Organisation_TRIPLE_Training_Toolkit	.docx	OA	CC-BY-4.0	Zenodo

	Name/Title of document	Format	Access type	Licence	Location
5	To_Do_TRIPLE_Training_Toolkit	.xlsx	OA	CC-BY-4.0	Zenodo
6	List_Past_Events_TRIPLE_Training_Toolkit	.docx	OA	CC-BY-4.0	Zenodo
7	Training_Objectives_Learning_Outcomes_TRIPLE_Training_Toolkit	.docx	OA	CC-BY-4.0	Zenodo
8	Internal_Training_Needs_Survey_TRIPLE_Training_Toolkit	.docx	OA	CC-BY-4.0	Zenodo
9a	Internal_Training_Needs_Results1_TRIPLE_Training_Toolkit	.png	OA	CC-BY-4.0	Zenodo
9b	Internal_Training_Needs_Results2_TRIPLE_Training_Toolkit	.png	OA	CC-BY-4.0	Zenodo
10	Post_Training_Survey_TRIPLE_Training_Toolkit	.docx	OA	CC-BY-4.0	Zenodo
11	Post_Training_Survey_Results_TRIPLE_Training_Toolkit	.xlsx	OA	CC-BY-4.0	Zenodo
12	Promotion_Dissemination_Template_TRIPLE_Training_Toolkit	.xlsx	OA	CC-BY-4.0	Zenodo
13	Enlarged_Audience_Template_TRIPLE_Training_Toolkit	.xlsx	OA	CC-BY-4.0	Zenodo

2. Overview of the outcomes of the training sessions

	Name/Title of document	Format	Volume	Access type	Licence	Location
1	Slides from the online training sessions	.pptx .pdf	1,33 GB	OA	CC-BY-4.0	Zenodo
2	Video recordings of the online training sessions	.mp4	17,5 GB	OA	CC-BY-4.0	NAKALA YouTube

FAIR data

This section describes the provisions made or envisioned to accommodate the FAIR principles and thereby future usage of the training resources.

Findable

All materials have been designed with sharing in mind. Following the “as closed as possible, as open as necessary” doctrine, all documents (see above) were produced to be published in Open Access in a public repository, Zenodo (affiliated to TRIPLE’s Zenodo collection) or Nakala.

To increase the findability of the materials, they are indexed on DARIAH Campus, a discovery platform for training materials related to digitally enabled arts and humanities.

Accessible and interoperable

Beyond making the training resources and their documentation available Open Access, a rich description of the events is available alongside the deposits, either as metadata or in README files. On the DARIAH Campus platform, training-specific metadata is added to the materials.

In terms of interoperability, community standards and formats (.pptx, .mp4) are adopted following the recommendations of the following, widely recognized paper: Ten simple rules for making training materials FAIR. PLOS Computational Biology, 16(5), e1007854. https://doi.org/10.1371/journal.pcbi.1007854

Reusable

As indicated in the overview table, the project is committed to make available outputs under a CC-BY 4.0 licence by default. This licensing policy is implemented in a way that is easily readable both for humans and for machines.

To further facilitate the reuse of both training and training support materials, open, editable formats are used for the deposits. All materials are contextualised with rich documentation and guides for reuse. A publication summarising the training design and delivery workflow will serve as a further context to enable clear accessibility and easy reusability of the documents.

Allocation of resources

The TRIPLE project will not have resources in place to curate and update the training resources after the conclusion of the project. However, this is aimed to be compensated with

rich documentation of the process and the outcomes. In addition, the date of the last update appears in each document.

The 6.3. Task force of the TRIPLE project has used this document as a project management tool to facilitate developing a common understanding and shared data management solutions across the project participants.

The costs for making the resources associated with T 6.3. FAIR are covered by the grant. During the project lifetime, Huma-Num and MWS are guiding the process that ultimately leads to the deposit of relevant components of the project in data repositories.

Data security

Task leaders are expected to have their back-up and storage policies in place followed by their own, local policies and back-up protocols using institutional cloud storage solutions (handled by their IT departments) and following the rules of GDPR.