


	 <p>Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration</p>
[07/2020]	Advancing Open Scholarship
	D2.1 – DATA ACQUISITION PLAN Version 1.0 – Draft PUBLIC
	H2020-INFRAEOSC-2019 Grant Agreement 863420

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

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

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



Data Acquisition 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	Data Acquisition and Categorization
WP Leader:	IBL-PAN
Deliverable identifier	D2.1
Contractual Delivery Date: 30/05/2020	Actual Delivery Date: 13/07/2020
Nature: Report	Version: 1.0 Draft
Dissemination level	PU/PR

Revision History

Version	Created/Modifier	Comments
0.0	Mélanie Bunel CNRS (HN), Laurent Capelli CNRS (HN), Arnaud Gingold CNRS (OE), Panayiota Polydoratou CNRS (OE)	Initial version
0.1	Mélanie Bunel CNRS (HN), Laurent Capelli CNRS (HN), Suzanne Dumouchel CNRS (HN) Arnaud Gingold CNRS (OE), Panayiota Polydoratou CNRS (OE), André Pacheco (COIMBRA), Maciej Maryl (IBL-PAN), Christopher Kittel (OKMAPS), Marta Błaszczczyńska (IBL-PAN), John Shepherdson (CESSDA), Twan Goosen (CLARIN)	Completeness and English revision
0.2	Maxi Schramm (OKMAPS), Giulio Andreini (Net7)	Peer review
0.3	Mélanie Bunel CNRS (HN), Laurent Capelli CNRS (HN), Arnaud Gingold CNRS (OE), Panayiota Polydoratou CNRS (OE), André Pacheco (COIMBRA), John Shepherdson (CESSDA), Christopher Kittel (OKMAPS), Maciej Maryl (IBL-PAN)	Completeness
0.4	Laurent Capelli CNRS (HN), Jean-Luc Minel (Paris-Nanterre University), Stéphane Pouyllau CNRS (HN), Mélanie Bunel CNRS (HN)	Internal review
0.5	Marta Błaszczczyńska (IBL-PAN), Arnaud Gingold CNRS (OE), John Shepherdson (CESSDA), André Pacheco (COIMBRA), Twan Goosen (CLARIN), Luca de Santis (Net7)	Completeness and English revision
1.0	Mélanie Bunel CNRS (HN), Laurent Capelli (HN), Suzanne Dumouchel (HN)	Final version

Acronyms

API	Application Program Interface
DC	Dublin Core
EOSC	European Open Science Cloud
FAIR	Findable Accessible Interoperable Reusable
HTML	HyperText Markup Language
IS	Innovative Services
JSON	JavaScript Object Notation
ML	Machine learning
OAI-PMH	Open Archives Initiative – Protocol for Metadata Harvesting
RDF	Resource Description Framework
SSH	Social Sciences and Humanities
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language

DRAFT

Abstract

This report describes the Data Acquisition Plan (DAP) with technical specifications to be implemented in order to collect metadata about the research outputs from Social Sciences and Humanities (SSH) in 9 languages¹ and to make them available through the future TRIPLE platform.

To reach this objective, the DAP, strategic step of the TRIPLE Core, defines the process of collecting metadata until their exposition in the TRIPLE database through a two-fold approach: 1) Metadata provision by processing chains of aggregation platforms and 2) Semantic enrichment and resource linking by the TRIPLE pipeline. A delivery platform will be the communication interface between both processes.

As a first phase, metadata are collected by aggregation platforms which are part of the consortium such as ISIDORE² or CESSDA³ (and others out of the consortium like OpenAIRE⁴, NARCIS⁵ etc.) and dropped on the delivery platform. To collect and expose their metadata, these platforms use generic processing chains called BUILD. In accordance with the TRIPLE recommendations and with their agreement, the BUILD chains will deliver selected metadata on a delivery platform, under the monitoring of the OPERAS⁶ Scientific Advisory Committee. This implies that the TRIPLE project creates a model, called TRIPLE data model, that the aggregation platforms might align with to be compliant with the TRIPLE platform. To start the project, the ISIDORE platform, developed by the coordinator⁷ of the TRIPLE project, had been chosen to be the first source of metadata, by using its processing chain “BUILD-I”, as indicated in the proposal. In the long run, to reach a satisfying level of exhaustivity, other BUILD chains will be added to cover the maximum of resources available in the whole SSH community worldwide.

In a second phase, by a connexion to the delivery platform, the TRIPLE pipeline will be able to collect, enrich and link the metadata corresponding to the 3 types of resources targeted by the project: 1) Research documents (publications and datasets), 2) Research projects and 3) Researcher profiles. The semantic enrichment will imply the creation of a TRIPLE thesaurus to align the vocabularies in the 9 languages.

The enriched and linked metadata will be then both stored in a tripleStore and indexed in the TRIPLE database and available through REST APIs for the Innovative Services (IS) to run their tools or for data providers to retrieve improved metadata.

¹ French, English, Spanish, Portuguese, Italian, German, Polish, Croatian, and Greek.

² <https://isidore.science/>

³ <https://www.cessda.eu/>

⁴ <https://www.OpenAIRE.eu/>

⁵ <https://www.narcis.nl/>

⁶ OPERAS coordinates services, practices and technology across the main actors in the SSH scholarly communication in Europe: <https://operas.hypotheses.org/>

⁷ The coordinator is the Research Infrastructure Huma-Num (operator): <https://www.huma-num.fr/>

Publishable Summary

The Data Acquisition Plan (D2.1) presents the main functionalities that the TRIPLE approach must cover, i.e. what it is supposed to do in order to satisfy the objectives of the project. More specifically, it provides a general description of the technical specifications in order to collect data in SSH disciplines to create the future TRIPLE platform.

This document is structured according to the following list:

- The first section presents the two-fold approach of the process including the metadata curation and the TRIPLE pipeline;
- The second section presents the targeted data sources;
- The third section describes the TRIPLE data model;
- The fourth gives an overview of the data acquisition roadmap.

The first section is a presentation of the two-fold approach, including 1) The metadata curation via BUILD processing chains interacting with 2) The TRIPLE pipeline thanks to a delivery platform. Firstly, it describes what the BUILD chains are composed of, from the metadata harvesting, allowing the connection to data repositories or webpages, to their dropping on the delivery platform with a focus on the BUILD-I chain. Then it explains briefly how the TRIPLE pipeline retrieves the metadata and injects them into an enrichment and linking process to make them available for the Innovative Services or other users.

The second section focuses on data sources and presents the types of data that are interesting for the project and explains where to find them to feed the future TRIPLE database. Also, it includes a presentation of the governance established in the data acquisition process.

The third section describes the TRIPLE data model based on the Schema.org⁸ ontology that the aggregation platforms might align with in order to be compliant with the TRIPLE platform. This TRIPLE data model considers the TRIPLE requirements, i.e. the Innovative Services needs and the FAIR principles⁹.

Finally, the data acquisition roadmap includes the description of the steps for the aggregation platforms to deliver their metadata on the delivery platform and a timeline between their extraction by the TRIPLE pipeline until their availability to the IS and users. Each step involves specified stakeholders.

⁸ <https://schema.org/>

⁹ <https://www.go-fair.org/>

Table of Contents

1 INTRODUCTION	7
2 OBJECTIVES OF THE DELIVERABLE	9
3 TRIPLE APPROACH	11
3.1 METADATA CURATION	12
3.1.1 BUILD CHAINS	12
3.1.2 BUILD-I CHAIN	13
3.2 TRIPLE PIPELINE	15
3.2.1 DELIVERY PLATFORM	15
3.2.2 TRIPLE SEMANTIC ENRICHMENT	16
3.2.3 RESOURCE LINKING	17
3.2.4 INDEXING AND EXPOSITION	17
4 METADATA SOURCES	19
4.1 TYPES OF RESOURCES	19
4.2 RESOURCES PROVIDERS	19
4.2.1 PUBLICATIONS AND DATASETS	19
4.2.2 RESEARCH PROJECTS	19
4.2.3 RESEARCHER PROFILES	20
4.3 RESOURCE MONITORING	20
5 TRIPLE MODEL	22
5.1 TRIPLE REQUIREMENTS	22
5.1.1 INNOVATIVE SERVICES (IS)	22
5.1.2 FAIRIFICATION PROCESS	23
5.2 TRIPLE DATA MODEL	23
5.2.1 SCHEMA.ORG	23
5.2.2 METADATA TERMS	24
5.2.3 BETA-MODEL	28
6 DATA ACQUISITION TIMELINE	29
7 CONCLUSION	31
8 GLOSSARY	32
9 REFERENCES	37

List of figures

<i>Figure 1. TRIPLE Platform Architecture with a focus on T2.1</i>	10
<i>Figure 2. Schematic and simplified representation of the two-fold approach to ingest metadata in the TRIPLE database</i>	11
<i>Figure 3. A generic processing chain</i>	13
<i>Figure 4. Example of a Sitemaps XML file</i>	14
<i>Figure 5. Enrichment process</i>	17
<i>Figure 6. detailed representation of the two-fold approach to ingest metadata in the TRIPLE database</i>	18
<i>Figure 7. TRIPLE data model and linking between the 3 types of resources. Legend: “CreativeWork” for research documents publications and datasets, “Project” for research projects and “Persons” for Researcher profiles.</i>	28
<i>Figure 8. Chronology to acquire metadata from the data providers until their exposition to users</i>	30

List of tables

<i>Table 1. List of minimum scientific, technical and ethical criteria for a data source to be accepted in the TRIPLE platform. first draft soon to be fleshed out and validated by the OPERAS SAC</i>	21
<i>Table 2. Metadata terms for publications and datasets (BETA-MODEL)</i>	25
<i>Table 3. Metadata terms for Research projects (BETA-MODEL)</i>	26
<i>Table 4. Metadata terms for Researcher profiles (BETA-MODEL)</i>	27

1 | INTRODUCTION

The TRIPLE platform will offer a deep discovery experience for future users by giving them the opportunity of exploring a network of experts and knowledge in SSH fields, creating collaborations and funding their projects.

The Data Acquisition Plan (DAP) is essential to the future of the TRIPLE platform as it will ensure the discovery of a variety of metadata collected from several sources. As a complement to the Data Management Plan (Deliverable 1.3), the DAP describes the process of data ingestion into the future TRIPLE database and will serve as a basis for the establishment of guidelines for future data providers (D2.2 “Data harvesting best practices document for data providers”). In the whole general architecture of the future TRIPLE platform (Figure 1), data acquisition is a part of the TRIPLE Core. This architecture shows that different services can work independently but also some of them are interdependent, and the database is one of the most important parts that will allow the other services to be set up.

In the case of TRIPLE, data acquisition means collecting metadata of scholarly outputs, research projects and researcher profiles, from SSH disciplines produced by BUILD processing chains in order to make them discoverable by the future users of TRIPLE. Metadata allow the description of database contents and are required to build and use systems that access and deliver data to user requests. Metadata aggregating by platforms is thus a foundational activity for the TRIPLE project. They are part of a cycle which starts with metadata creation by the providers, like research centres, libraries, museums, etc. and stored into repositories until their harvesting, indexing and exposition by external databases for reuse. The exploration of existing metadata in the world of SSH reveals that TRIPLE faces three main challenges: 1) Heterogeneity of data, 2) Heterogeneity of sources and 3) Open data access and availability. When heterogeneous metadata are brought together in a single system, their formalizations must be homogenized and integrated in order to support the delivery system. To overcome these challenges, TRIPLE is building an ambitious platform architecture of interoperability supporting this heterogeneous environment to produce high quality metadata¹⁰ based on metadata aggregators. The level of metadata quality will directly affect the visibility and the reuse of the resources by the users of the discovery platform. As expressed in the proposal, this project’s ambition starts with building a data acquisition plan in order to reach the platform specifications including:

- 1) The implementation of the Innovative Services (IS)

The Innovative Services are applications and tools that are not part of the core of the TRIPLE platform. These applications and tools will work on top of this core and deliver additional services for SSH researchers and other TRIPLE stakeholders. Well documented APIs will make it easy to integrate additional tools as services or data providers in the platform.

It includes at this stage:

¹⁰ A metadata record is qualitative when it reaches accuracy, provenance, consistency and logical coherence, timeliness, accessibility, subject term specificity and exhaustivity.

- A recommender system¹¹;
- A visualization tool¹²;
- Open annotation tool¹³;
- Trust Building System (TBS)¹⁴;
- A crowdfunding system.

2) Using data from European Open Science Cloud (EOSC¹⁵)

TRIPLE should be able to use EOSC data such as those provided by OpenAIRE, ISIDORE, NARCIS, MOSA¹⁶, Europeana¹⁷ and so on. These aggregation platforms, compliant with FAIR principles, provide services to help the research community to adopt and make the transition to the Open Sciences easily, like resources, guidelines and training. Being FAIR is an objective and the road to reach this requirement will be done by developing a solution with basic steps to progress in the level of FAIRification.

DRAFT

¹¹ User interaction data will be tracked to form the basis for the project's personalized services. Such services improve the user experience, support decision making and assist in the finding of relevant items and peers.

¹² Creation of knowledge map representations for the specific purpose of visual presentation of TRIPLE search results.

¹³ Creation of digital annotations, i.e., marginalia on digital resources.

¹⁴ Referral system designed as a new generation of social network informed by collective intelligence techniques, complexity theory, and social sciences. It aims to provide connectivity without sacrificing trust in order to enable multi-stakeholder cooperation.

¹⁵ <https://www.eosc-portal.eu/>

¹⁶ <https://www.mosa-research.be>

¹⁷ <https://www.europeana.eu/>

2 | OBJECTIVES OF THE DELIVERABLE

The objective of this deliverable is to provide the effective metadata collecting ecosystem for the TRIPLE platform. The DAP defines the general technical specifications for this process which includes the definition and harmonisation of the metadata. Further to this, a support will be provided to TRIPLE partners, OPERAS community and other potential providers, and partnership with aggregation platforms, especially from the CO-OPERAS IN¹⁸ community and the D2.2, in order to comply with the platform specifications. This support will materialise through various activities that will ensure the SSH community engagement and may include (but will not be limited to) workshops, tutorials, webinars, guidelines, documentation and, where needed, interactive communications.

The DAP, the establishment of guidelines and the support provided to the data providers, as WP2 tasks, are coordinated by IBL PAN until M30. Between M30 and M42, this work will be supervised by the Technical Board of TRIPLE, composed of technical WP leaders, with the help of the OPERAS Scientific Advisory Committee (SAC)" (operational in September 2020). The TRIPLE data acquisition ecosystem will be documented in living documents in order to facilitate the work of OPERAS SAC and provide clear information to the new data providers and aggregation platforms.

DRAFT

¹⁸ <https://operas.hypotheses.org/2697>

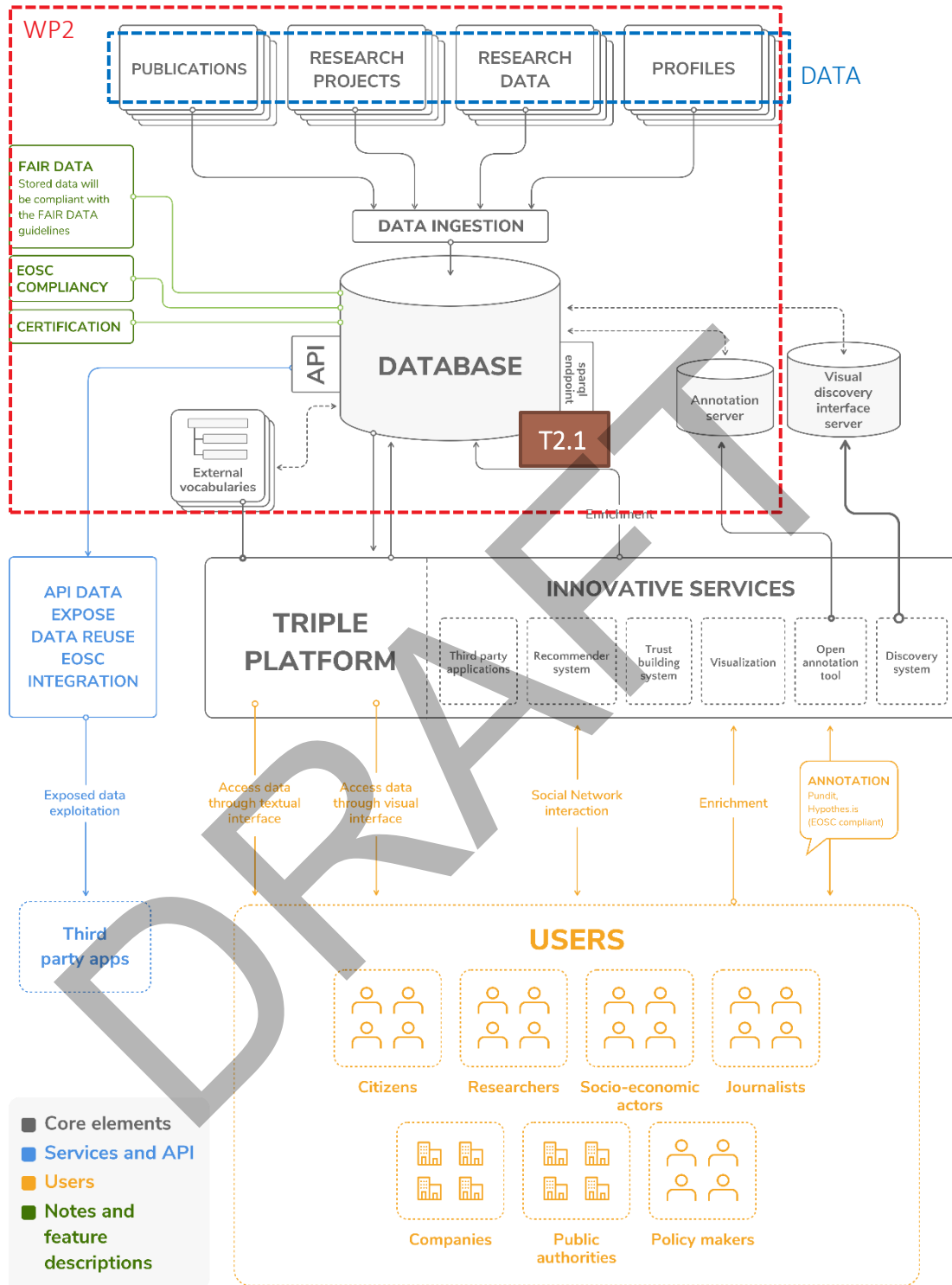


FIGURE 1. TRIPLE PLATFORM ARCHITECTURE WITH A FOCUS ON T2.1

3 | TRIPLE APPROACH

Data acquisition is included in a two-fold approach (Figure 2): 1) Metadata curation by BUILD processing chains of aggregation platforms (National or European) and 2) The TRIPLE pipeline focused on the semantic enrichment and resource linking.

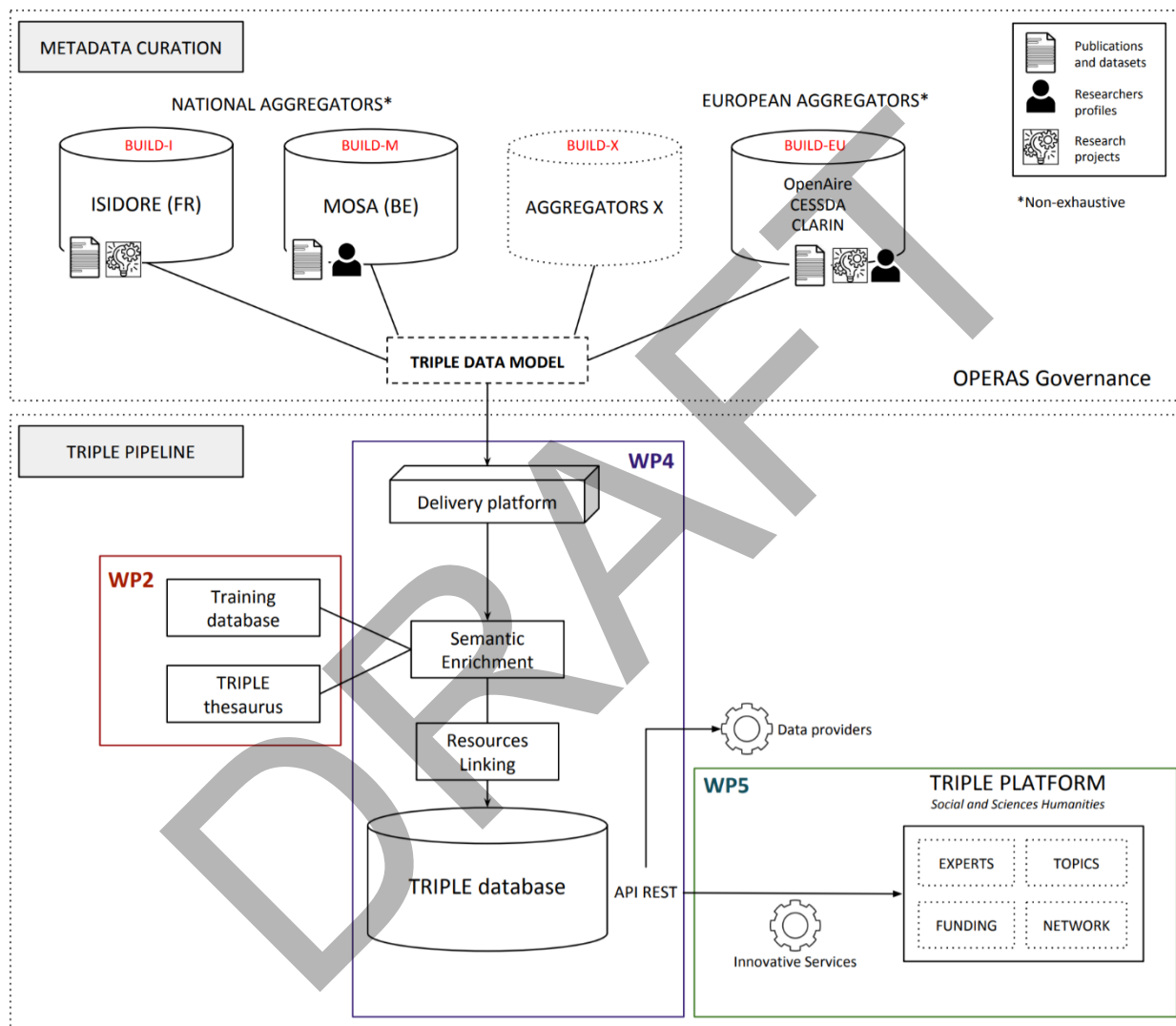


FIGURE 2. SCHEMATIC AND SIMPLIFIED REPRESENTATION OF THE TWO-FOLD APPROACH TO INGEST METADATA IN THE TRIPLE DATABASE

It is assumed that TRIPLE leverages metadata harvested by existing aggregation platforms. More precisely, TRIPLE relies initially on the processing chains, called BUILD, and as a first step by using the BUILD-I chain, currently operational and provided by ISIDORE platform. In the future of the project, TRIPLE aims to exploit other processing chains called BUILD-X (for any platform compatible with the recommendations given by TRIPLE). The interest of this conceptual choice is to avoid developing a platform with a specific processing chain when there are already many of

them and when these platforms have strongly capitalized their experience. To be compliant with TRIPLE, these BUILD chains will be invited to align their own model, using specific metadata standards and ontologies, on the TRIPLE data model developed by WP2 “Data acquisition and categorisation”. The quality of delivered metadata is under the monitoring of the task 2.2 “Individual support to data providers”. The BUILD chains communicate with the TRIPLE pipeline by dropping their metadata packets on a delivery platform developed by the WP4 “Integration and building of TRIPLE platform”.

In a second phase, the TRIPLE pipeline retrieves these packets and proceeds of a semantic enrichment of collected metadata and creates a linking between them, which is considered as the capital gain of this platform and currently non-existent at this requirement level in the landscape of the discovery platforms. This phase is introduced in this present deliverable and will be detailed in the D2.3 “Report on machine learning” and D2.4 “Report on creation and mapping vocabularies”. Finally, the TRIPLE pipeline integrates metadata in an RDF tripleStore and then indexes the metadata into the TRIPLE search engine. REST APIs allow the Innovative Services IS to set up their tools or to any users or data providers to retrieve the metadata.

3.1 Metadata curation

3.1.1 Build chains

A BUILD chain is the name for a processing chain created by an aggregating platform. Generically, it is composed of 3 steps (Figure 3): 1) Metadata harvesting 2) Semantic enrichment and 3) Indexing in a database. The BUILD chains framework and content are under the responsibility of aggregation platforms which process the metadata they harvest from data providers. Also, according to the TRIPLE recommendations (see section 5), the aggregation platforms are responsible for the selection of metadata packets they decide to deposit to the delivery platform. This selection is called scientific, disciplinary and linguistic curations.

There are 3 types of resources that the BUILD chains might deliver and are interesting for TRIPLE:

- Metadata describing scholarly publications and research datasets in compliance with the TRIPLE data model for Documents (see section 5).
- Metadata describing research projects in compliance with the TRIPLE data model for Projects (see section 5).
- Metadata describing researcher profiles in compliance with the TRIPLE data model for Persons (see section 5).

As a first step, TRIPLE will rely on the BUILD-ISIDORE (BUILD-I) chain to curate resources (Publications, datasets, projects and profiles) to feed the delivery platform and to be able to set up the enrichment process and the Innovative Services. The BUILD-I chain is developed and maintained by Huma-Num as part of the ISIDORE platform.

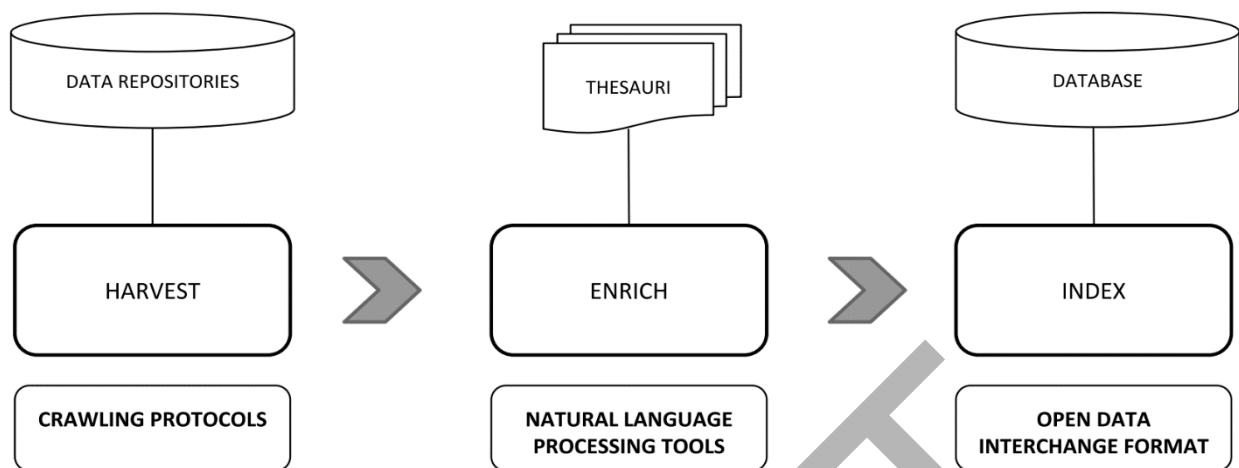


FIGURE 3. A GENERIC PROCESSING CHAIN

3.1.2 BUILD-I chain

ISIDORE, via a BUILD-I chain, ISIDORE is a French platform which collects and indexes digital research data in the fields of humanities and social sciences. It promotes access to open access data produced by research and higher education organizations, laboratories, libraries and research teams.

A. BUILD-I HARVESTING PROCESS

ISIDORE harvests metadata and full text (when available) from publications, corpora, databases and scientific news, accessible on the web and in open and interoperable standards. ISIDORE harvests metadata in 3 languages: French (produced in France or in the French-speaking world), English and Spanish.

The harvesting process is a mechanism that allows metadata to be collected from a remote catalogue (or a remote database) and stored in a local space (server), called a repository, for faster access. This harvesting process is done by ISIDORE regularly and launched manually by the developers, at least once a month to get the last updates. To perform this harvesting, both the data providers and the harvester must use the same technical protocol: the BUILD-I chain harvests from open archives using the OAI-PMH Version 2 protocol and on the Web through HTML embedded metadata (RDFa) listing by a Sitemaps file.

OAI-PMH protocol

The OAI-PMH protocol was developed in 1999 as part of the Open Archives Initiative¹⁹. Originally intended to allow exchanges between open archives and documentary portals, the OAI-PMH protocol has been quickly adopted for other uses, in particular in the field of heritage and digital libraries. It allows data providers to expose their metadata on the Web. It is an "overlay" to the standard web protocol HTTP, defining six specific query verbs: GetRecord, Identify,

¹⁹ <https://www.openarchives.org/pmh/>

ListIdentifiers, ListMetadataFormats, ListRecords and ListSets. The OAI-PMH protocol implies 2 stakeholders: 1) The data provider, who creates structural metadata and exposes them in a repository for harvesting and 2) The service provider, who harvests and enriches the structured metadata, and provides a searchable interface to retrieve metadata records.

An OAI-PMH repository has two levels of granularity:

- The SET corresponds to a coherent set of resources whose scope is the responsibility of the provider. It defines a hierarchy of sets with an inheritance mechanism, by specifying in the name of the set the name of the parent and child sets separated by the character ":";
- The RECORD gathers all the metadata of a resource in the manner of a bibliographic record, corresponding to an indexed document. A record must be expressed according to an application profile (vocabulary allowing the description of content). By default, the OAI-PMH protocol uses Dublin Core²⁰ to describe the scientific information it disseminates in a root element: oai_dc. In addition to this description in Dublin Core, each record can be described using one or more alternative metadata standards.

The BUILD-I harvesting process uses the combination of both OAI-PMH ListIdentifiers and GetRecord verbs from the OAI-PMH protocol. Also, the “identifier” value is used in the ISIDORE database as the unique identifier²¹. The delete option of the OAI protocol is also considered by the harvesting process. To update²² the records, the harvesting process will be set up to harvest the repositories either in a full mode (the whole repositories) or in an incremental mode (only the last modifications) by using OAI-PMH “from” argument. In order to inform the BUILD-I chain about modifications on records, data providers must change the “datestamp” value in the “from” argument in the OAI records. Also, the ISIDORE bot is also configured to accept or reject some OAI sets.

Sitemaps protocol

Initiated by Google, the Sitemaps protocol indicates to search engines the resources of a website to be indexed. It assumes the form of an XML file which contains, for each resource, its URL, the date of its last modification, the frequency of revision and the relative importance compared to the other URLs of the site (Figure 4).

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.com/</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
</urlset>
```

FIGURE 4. EXAMPLE OF A SITEMAPS XML FILE

²⁰ <https://www.dublincore.org/>

²¹ “A unique identifier unambiguously identifies an item within a repository. The unique identifier is used in OAI-PMH requests for extracting metadata from the item. Items may contain metadata in multiple formats. The unique identifier maps to the item, and all possible records available from a single item share the same unique identifier.” Source: <https://www.openarchives.org/OAI/openarchivesprotocol.html#UniqueIdentifier>

²² Some records can be added, modified or deleted.

In the context of the BUILD-I harvesting process, the use of the Sitemaps protocol will guide the collection of web pages and limit it to the most relevant pages, according to the project objectives. Thus, it will allow:

- The exhaustive indexing of a resource when the structure would be too complex to present in an OAI-PMH repository (for example a monograph or an edition of sources);
- Support for situations in which setting up an OAI-PMH repository would be too complex to harvest.

Therefore, the providers could expose two Sitemaps files: 1) One for traditional search engines which would reference all the web pages of the site and 2) A second for the BUILD-I harvester which would be limited to pages with relevant scientific content determined by the Scientific Committee. In this second case, the provider will indicate the URL of the Sitemaps feed intended for TRIPLE. This URL is the unique identifier for Sitemap protocol written in the sitemap XML document. The “Lastmod²³” field must be used to indicate to the TRIPLE engine new modifications in the records.

As a Sitemaps schema does not allow the descriptive metadata of the resource to be expressed directly in the XML stream that composes it, the resource providers must integrate these metadata, according to RDFa syntax, within the same HTML page that they want to be indexed by a BUILD processing chain. ISIDORE retrieves RDFa metadata in the DC and DCTerms standards.

B. BUILD-I ENRICHMENT PROCESS

Once harvested, the BUILD-I chain is able to enriched metadata by using three Natural Language Processing (NLP) tools: categorization, normalization and annotation. These tools are available in the NLP toolbox called “ISIDORE On Demand²⁴”. It is a set of semantic enrichment tools accessible through APIs exposed by Huma-Num. Metadata are enriched in three languages (English, Spanish and French) by cross-referencing with thesauri produced either by the scientific community (GeoEthno, Pactols...) or by major institutions in the field of higher education and research (Rameau, LCSH, BNE, Gemet, Lexvo, GeoNames...).

C. BUILD-I INDEXING

The Indexing is carried out by the ISIDORE search engine then stored in a database RDF-compliant for all structured data (original metadata and enrichment result).

3.2 TRIPLE pipeline

3.2.1 Delivery platform

²³ Last modifications.

²⁴ <https://rd.isidore.science/ondemand/en/>. To be differentiated from the ISIDORE platform.

The BUILD processing chains produce metadata adapted to the TRIPLE format that they can deposit on a delivery platform at regular periods to be defined in negotiation with platform stakeholders. This delivery platform is a communication interface between the BUILD-X chains and the TRIPLE pipeline. Processing chains wishing to deliver their metadata to TRIPLE, will be able to carry out a selection phase of their metadata on scientific, disciplinary, linguistic and legal criteria, called “scientific curation” in accordance with the recommendations of TRIPLE and following the criteria provided by OPERAS SAC. On this delivery platform, the metadata packets are exchanged on the principle of the push-pull model. In this process, the interaction between a supplier (a BUILD-X chain) and the event channel is “push”, and between a consumer (TRIPLE) and the event channel is “pull”. A supplier generates an event and passes it to the event channel. The channel does not transfer the event to registered consumers until consumers pull for the event data. In this process, both suppliers and consumers actively initiate the interaction with the channel and the event channel acts as a queue component. If suppliers and consumers operate at different rates, the channel may need an unbounded queue to store the events. If this is not practical, the channel may need to implement some dropping mechanism or blocking when the queue is full.

On this delivery platform, TRIPLE services are capable of:

- Retrieving the resources from the BUILD-X chains and extract them with a software components “Extractor”;
- Checking the conformity of the metadata with the TRIPLE data model;
- Preparing to ingest metadata into the TRIPLE storage system;
- Sending a signal (FLAG-TRIPLE) on the event channel to indicate that the resources have been loaded.

3.2.2 TRIPLE semantic enrichment

During this step, the TRIPLE pipeline proceeds to its own multilingual semantic enrichment of the metadata comprising 3 types of enrichment (Figure 5):

- Normalization of metadata values;
- Classification by machine learning (ML): automatic categorization of the indexing document using the MORESS disciplines²⁵ in the 9 languages foreseen by the project;
- Semantic annotations using a multilingual TRIPLE thesaurus, in the 9 languages foreseen by the project, firstly based on LCSH thesaurus with the aim to include 3000 concepts and refining with other existing thesauri. It includes a disambiguation tool²⁶ using WIKIDATA. This TRIPLE thesaurus is scalable and will be maintained by a dedicated team. The update and freshness of the thesaurus is absolutely essential to match with the concept of knowledge discovery, as one of the objectives of TRIPLE.

²⁵ European University Association, “MORESS-Mapping of Research in European Social Sciences and Humanities”, Final report, 2006. The MORESS project identifies 27 disciplines in the SSH fields. They are currently used by Isidore and the French repository HAL-SHS (<https://halshs.archives-ouvertes.fr/browse/domain>).

²⁶ NERD: Name Entity Recognition and Disambiguation.

These different steps of enrichment are allowed by using the NLP toolbox “ISIDORE On Demand”, integrating categorization, normalization and annotation. TRIPLE relies on this toolbox as a customer of Huma-Num, to enrich, categorize and annotate the TRIPLE metadata in the defined 9 languages. The tools will be trained with the several thesauri developed for TRIPLE and specifically by the tasks 2.3 and 2.4.

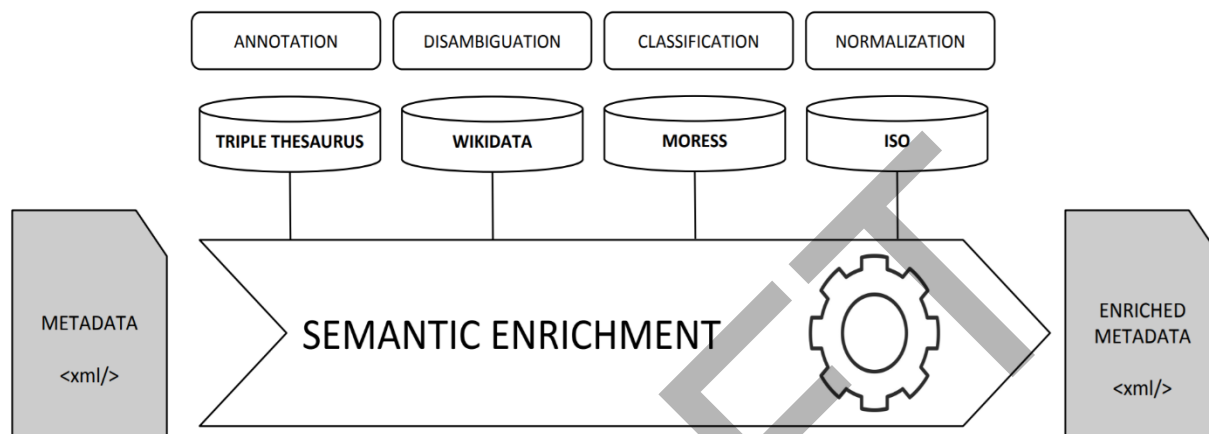


FIGURE 5. ENRICHMENT PROCESS

3.2.3 Resource linking

One of the objectives of the project, also constituting one of the added values of the platform, is the linking between the different resources extracted on the delivery platform. This linking will be realized by algorithms developed by WP4.

3.2.4 Indexing and exposition

Once the metadata has been enriched and linked, they are both stored in an RDF tripleStore and indexed by a search engine in the database. This indexing will be done using the ElasticSearch²⁷ indexing tool. The metadata are then available through REST APIs for reuse by the Innovative Services and data providers for recovering enrichment on their metadata.

The following scheme (Figure 6) provides a detailed description of the whole TRIPLE approach to acquire and process metadata.

²⁷ <https://www.elastic.co/elasticsearch/>

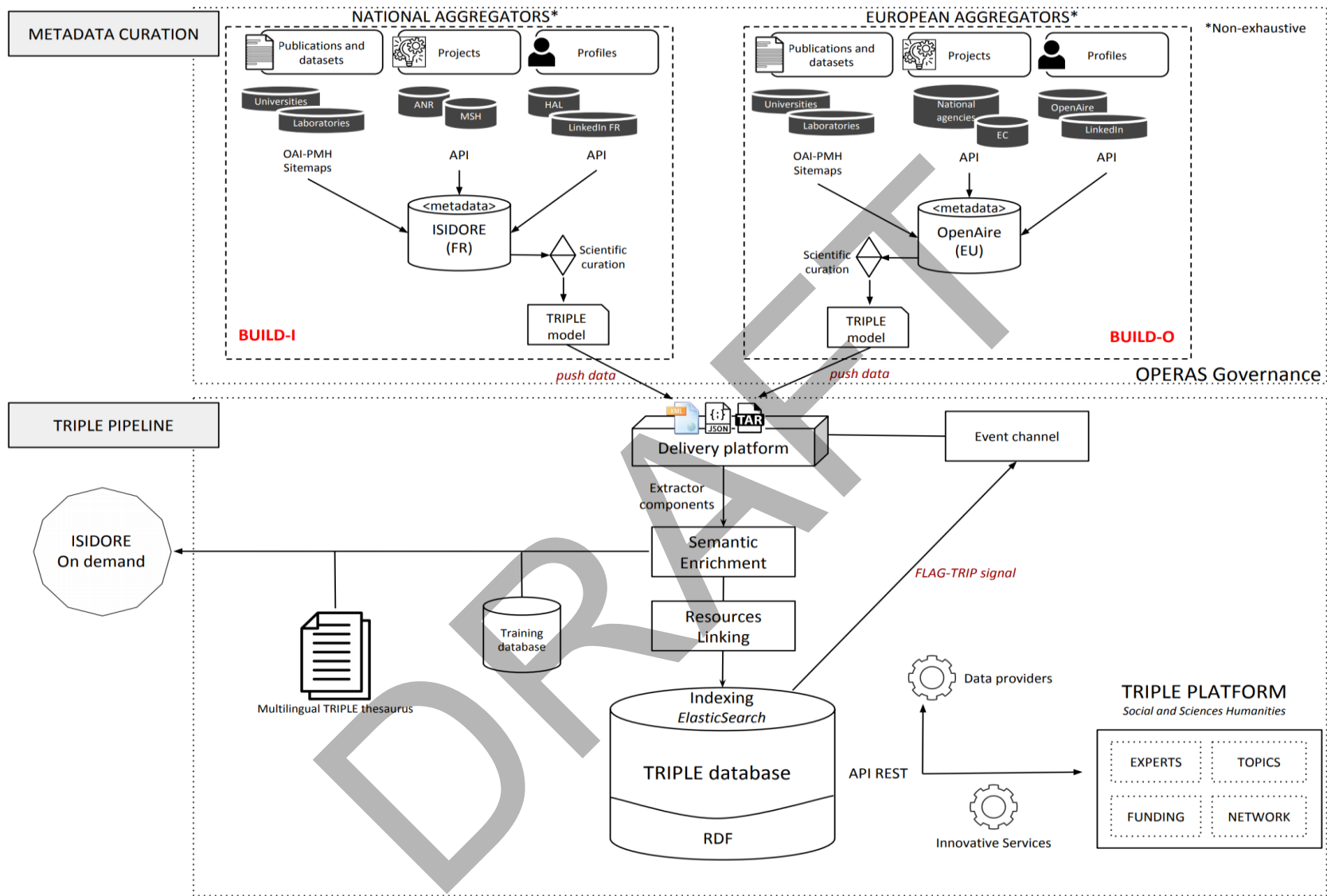


FIGURE 6. DETAILED REPRESENTATION OF THE TWO-FOLD APPROACH TO INGEST METADATA IN THE TRIPLE DATABASE.

4| METADATA SOURCES

4.1 Types of resources

The harvesting process of BUILD-X processing chains will focus on metadata related to:

- Publications, as in the documents produced by the scientific community (like scholarly articles, books, thesis, blog posts, images, and any other research reports);
- Research data or datasets, as in the data produced by the researchers which are raw data (like survey data for example);
- Research projects, as existing research projects in SSH fields funded by the European Commission, national funding agencies, crowdfunding or any other funded or non-funded projects;
- Researcher profiles. A BUILD-X processing chain provides the possibility of semi-automatically retrieving person's information through their ORCID accounts. A similar solution could be used with other researcher profile platforms like ResearchGate or LinkedIn.

4.2 Resources providers

A BUILD-X chain can provide either one type of resources or several types of resources.

4.2.1 Publications and datasets

Metadata on publications and datasets will firstly be provided by the BUILD-I chain as written previously. The BUILD-I chain contains a large stock of metadata about research documents and will be their main provider in the TRIPLE database. At this stage, the other aggregation platforms targeted by TRIPLE which might deliver metadata are:

- European platforms like CLARIN and CESSDA especially for datasets;
- The European platform OpenAIRE for its European coverage exhaustiveness;
- The European platform Europeana especially for its richness on Cultural Heritage;
- National platforms like NARCIS (Netherlands) or MOSA (Belgium) etc.

In order to localize other data sources, through the T2.2 "Individual support to data providers", TRIPLE will realize a cartography of existing data providers which might be likely to feed the TRIPLE database and increase its coverage. If these data providers are not already connected to an aggregation platform, especially for small ones, on their request, they will be invited to join either an existing National platform or ISIDORE or OpenAIRE to appear in the TRIPLE platform.

4.2.2 Research projects

Research projects will also be retrievable in the platform. The inputs will come from:

- European platforms like OpenAIRE, CORDIS²⁸, CLARIN²⁹, CESSDA, Europeana;
- National funding agencies databases;
- Small non-funding projects conducted in any laboratories, or by independent researchers;
- The crowdfunding tool.

4.2.3 Researcher profiles

TRIPLE will allow us to discover experts in SSH fields. These profiles will come from:

- Aggregation platforms which collect profiles or allow to create profiles on their platforms;
- Professional network platforms;
- ORCID registry.

4.3 Resource monitoring

The OPERAS Scientific Advisory Committee “OPERAS SAC” will be responsible for validating a list of minimum scientific, technical and ethical criteria (Table 1) which will determine whether or not new sources will be added to the search engine. OPERAS SAC will be established in September 2020 and will be responsible for validating this list by December 2020. At the time of submitting this deliverable OPERAS was collecting declarations of interests to become SAC members through an open call. A dedicated technical team will be responsible for accepting or not a source. However, if there is any doubt, OPERAS SAC can be solicited for decision. It means that, occasionally, for ambiguous sources, it will be possible to request OPERAS SAC to determine whether they should be included into the discovery platform.

²⁸ “The Community Research and Development Information Service (CORDIS) is the European Commission's primary source of results from the projects funded by the EU's framework programmes for research and innovation (FP1 to Horizon 2020)”. Source: <https://cordis.europa.eu/about/en>

²⁹ <https://www.clarin.eu/>

TABLE 1. LIST OF MINIMUM SCIENTIFIC, TECHNICAL AND ETHICAL CRITERIA FOR A DATA SOURCE TO BE ACCEPTED IN THE TRIPLE PLATFORM. FIRST DRAFT SOON TO BE FLESHED OUT AND VALIDATED BY THE OPERAS SAC.

LEVEL OF CRITERIA	MINIMAL REQUIREMENTS
SCIENTIFIC	Open Access data
	Social and Scientific Humanities scope
TECHNICAL	TRIPLE data model compliancy
ETHICAL	GDPR ³⁰ (see the Data Management Plan)

DRAFT

³⁰ <https://gdpr-info.eu/>

5 | TRIPLE MODEL

In order to increase the visibility of their data, repositories expose them, very often, through more than one metadata standard (see Glossary), which can be schemas, vocabularies, serialization, etc. At the same time, the metadata standards are closely related, for strategic reasons, to renowned³¹ aggregation platforms requirements, through either total or partial alignment. So, the use of metadata standards by the providers is made according to their internal needs, on the one hand, and to the harvesters' requirements, on the other.

One of the challenges of TRIPLE is to deal with the heterogeneity of metadata provided by the BUILD-X chains, which defines the available information. To be in line with the TRIPLE requirements essential to reach the objectives of the project, TRIPLE is establishing a *lingua franca* for the platform, the TRIPLE data model. By the creation of partnerships with TRIPLE, the BUILD-X chains will align their own metadata model with the TRIPLE model to deliver their metadata on the delivery platform.

It will be the mission of T2.2, with a dedicated Data steward, to support the providers that they can provide richer metadata to their harvesting platform in order to be enriched (and to appear) in the TRIPLE platform, to benefit from the IS like visualization or annotations and also from the whole discovery services. This close collaboration will allow harvesting of high-quality metadata and thus increase the discoverability of the data providers' content. The T2.2 might also provide support to the aggregation platform to reach the TRIPLE data model.

5.1 TRIPLE recommendations

In order to provide a high discovery experience, the TRIPLE data model must be built according two requirements:

- Allow the implementation of the Innovative Services;
- Being a reference in terms of FAIR data.

5.1.1 Innovative Services (IS)

The Innovative Services rely on the harvesting process and the semantic enrichment to be able to work, as they depend on the quality of harvested metadata.

Both the discovery system (visualization) and the Recommender System are particularly concerned by metadata and more specifically by having at least:

- The title of the publication or the title of the study for datasets;
- The date of publication or datasets creation;
- The language of the resource;
- The description of the resource including the abstract;
- The subject of the resource;

³¹ Among Open Access renowned harvesters, we can find BASE, CORE, Europeana, OpenAIRE, CLARIN or ZENODO.

- The file location with access rights.

5.1.2 FAIRification process

The TRIPLE data model is built according to the FAIR principles and will therefore facilitate their adoption by a wider community. More specifically, the Data Acquisition Plan will ensure to match the following principles:

- Findable:
 - Persistent identifiers, either retrieved or generated;
 - Rich metadata, thanks to the TRIPLE data model;
 - Direct link to the full resource;
- Accessible:
 - Use of the open and free protocol OAI-PMH.
- Interoperable:
 - Use of the standard knowledge representations;
 - Alignment with aggregators guidelines.
- Reusable:
 - Indexing of open access content;
 - Clear information on access rights for sensible data;
 - Implementations conducted in other tasks will imply further FAIRification of TRIPLE contents;
 - The use of domain-relevant vocabularies, such as RAMEAU, LCSH;
 - The content representation available in RDF triplets and a SPARQL endpoint.

5.2 TRIPLE data model

5.2.1 Schema.org

The proposal proposes "*Metadata records produced by TRIPLE will be published using the following standard vocabularies: Component Metadata Infrastructure, Dublin Core Metadata Element Set and DCMI Metadata Terms.*"[11]. Considering that these descriptions are limited in their semantic expressivity, we suggest to specify a TRIPLE data model based on the Schema.org ontology. It is an ontology whose scope (classes, entities) is much stronger than any other metadata standards. On the one hand, the semantic of classes and properties is more accurate and detailed than other standards (for example Dublin Core). It is also in constant evolution and follows the evolution of the web. The W3C community makes Schema.org evolve according to new needs. This ontology is the most suitable to describe the different types of resources that TRIPLE targets. The TRIPLE data model will be a recommendation for scientific information aggregation platforms

(such as ISIDORE, OpenAIRE, NARCIS, etc.) which will have to deliver their metadata packets according to this model.

Thanks to a partnership with the aggregators, the TRIPLE data model will be implemented by the BUILD-X chains by a crosswalk and a mapping process with their own data model with the help of a style sheet provided by TRIPLE. The proposed TRIPLE data model is a beta-version. By September 2020, an advanced TRIPLE data model will be available. As the TRIPLE platform will be an evolutive platform, this TRIPLE data model is evolutive as well. The adjustments will happen as progress of technical development and tests. It will result in the establishment of specifications in a living document regularly updated all along the project with several refinements. The first specifications are described below.

5.2.2 Metadata terms

The TRIPLE data model is built with metadata elements for 1) Publications and research datasets in Table 2., 2) Research projects in Table 3., and 3) Researcher profiles in Table 4. For the publications and datasets, minimal metadata elements have been identified considering what TRIPLE would try to obtain from the data providers as minimum descriptive metadata to be harvested by the aggregators. The extended metadata elements have been identified as other descriptive metadata to harvest having taken into consideration TRIPLE objectives in an ideal case, i.e., satisfy the Innovative Services' needs, and follow the FAIR principles. These metadata will be enriched by other metadata during the semantic enrichment process. At this stage, the proposed model is a Beta-version. Also, a whole ontology is currently in progress to structure the metadata and their relations. This ontology aims at specifying as completely as possible the semantics of the different fields that describe resources delivered as packets by BUILD chains but does not presume in any way how these packets will be implemented.

TABLE 2. METADATA TERMS FOR PUBLICATIONS AND DATASETS (BETA-MODEL). *MINIMAL ELEMENTS

METADATA TERMS	DESCRIPTION	ELEMENTS	CARDINALITY
Title*	Title of the publication	schema:headline	1...N
Landing Page	Link to the landing page of the document	schema:mainEntityOfPage	1...N
UrlDigitalDocument	Link to the full resource (URL) with AccessRights (not necessarily OpenAccess)	schema:url	1...N
Identifier*	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system	schema:identifier	1...N
Date*	Date of publication/creation	schema:datePublished	1...N
Language	Language of the content of the resource	schema:inLanguage	1...N
Creator*	Author of the publication/datasets	schema:creator	1...N
Contributor	Entity responsible for making contributions to the content of the resource (person, organization)	schema:contributor	0...N
Publisher	Entity responsible for making the resource available	schema:publisher	1...N
Publication type	Type of document	rdfs:type	1...N
Keywords	Keywords or tags used to describe this content. Multiple entries in a keywords list are typically delimited by commas.	schema:keywords	1...N
Description =Abstract	Textual description of the content	schema:description	0...N
Format	Specify the format of the resource	schema:encodingFormat	0...N
Source	A reference to a resource from which the present resource is derived.	schema:isBaseOnURL	0...N
Funding reference	Information on project funding	schema:funder	0...N
Coverage	Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity)	schema:temporalCoverage schema: spatialCoverage	0...N
Audience	For whom the resource is useful	schema:audienceType	0...N
License	A license document that applies to this content, typically indicated by URL.	schema:license	0...1

TABLE 3. METADATA TERMS FOR RESEARCH PROJECTS (BETA-MODEL)

METADATA TERMS	DESCRIPTION	ELEMETS	CARDINALITY
Title	Title of the project	schema:name	1...N
Identifier	The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links.	schema:identifier	1...N
Start Date	Date of the beginning of the project	schema:startDate	1
End Date	Date of the end of the project	schema:endDate	1
Duration of the project	The duration of the item (movie, audio recording, event, etc.) in ISO 8601 date format.	schema:duration	1
Funding agency	Funder of the project	schema:funder	0...N
Funding type	Type of funding: Granted or Crowdfunded	schema:fundingScheme	0...N
Coordinator	Coordinator of the project	schema:creator	0...N
Organization	Organization which coordinates or participates to the project	schema:affiliation	0...N
Description	Textual description of the project	schema:description	0...N
Link to the project	Link to the project	schema:URL	1...N

TABLE 4. METADATA TERMS FOR RESEARCHER PROFILES (BETA-MODEL)

METADATA TERMS	DESCRIPTION	ELEMENTS	CARDINALITY
Last Name	Last Name of the person	schema:familyName	1...N
First Name	First name of the person	schema: givenName	1...N
Identifier number	ORCID, other	schema:identifier	0...N
Affiliation	An organization that this person is affiliated with. For example, a school/university, a club, or a team.	schema:affiliation	1...N
Affiliation link	Link to the organization where the person is currently working	schema:url	1...N
Nationality	Nationality of the person.	schema:nationality	1...N
Occupation	Current occupation of the person	schema: hasOccupation	1...N
Date	Start Date of the current occupation	schema:startDate	1...N
Education	The level in terms of progression through an educational or training context. Examples of educational levels include 'beginner', 'intermediate' or 'advanced', and formal sets of level indicators.	schema:educationLevel	1...N
School	University or school where the education level was obtained	schema:name	1...N
Date of degree	Date when the degree was obtained	schema:dateIssued	1...N
Skills/expertise	Of a Person, and less typically of an Organization, to indicate a topic that is known about - suggesting possible expertise but not implying it. We do not distinguish skill levels here, or relate this to educational content, events, objectives or JobPosting descriptions.	schema:knowsAbout	1...N
Website	Website of the person	schema:url	1...N

5.2.3 Beta-model

The TRIPLE platform aims to link the three delivered and enriched resources. The following scheme (Figure 7) represents how these resources will be linked:

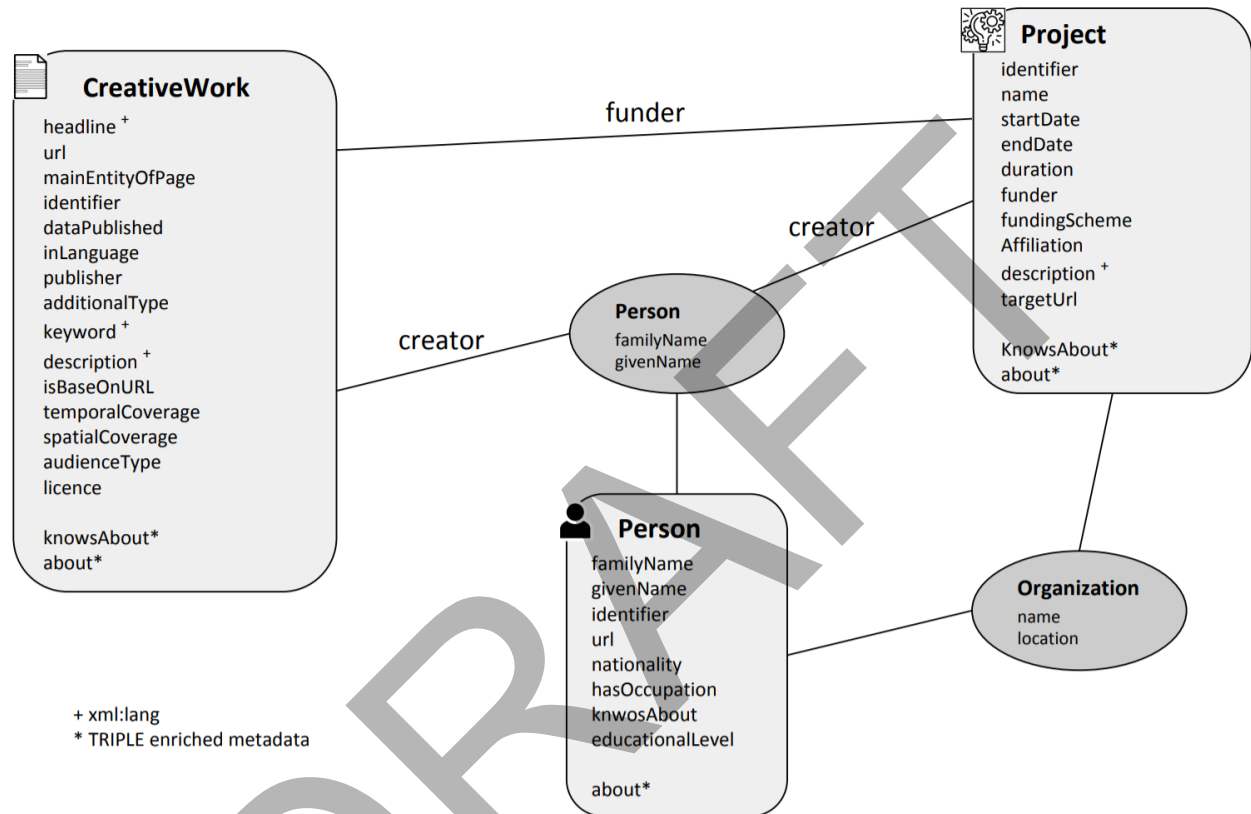


FIGURE 7. TRIPLE DATA MODEL AND LINKING BETWEEN THE 3 TYPES OF RESOURCES. LEGEND: "CREATIVEWORK" FOR RESEARCH DOCUMENTS PUBLICATIONS AND DATASETS, "PROJECT" FOR RESEARCH PROJECTS AND "PERSONS" FOR RESEARCHER PROFILES.

6 | DATA ACQUISITION TIMELINE

The Data Acquisition Plan is the mechanism for feeding the TRIPLE database. It will start with the contact with the data providers and the aggregation platforms. It involves several stakeholders during a defined timeline. The following scheme (Figure 8) describes the chronology to acquire metadata and to make them available to the users at the end of the TRIPLE pipeline. The metadata curation phase is out of control of TRIPLE in terms of time. It is possible to estimate the timeline from the moment when the resources are dropped on the delivery platform, but not before this step.

DRAFT

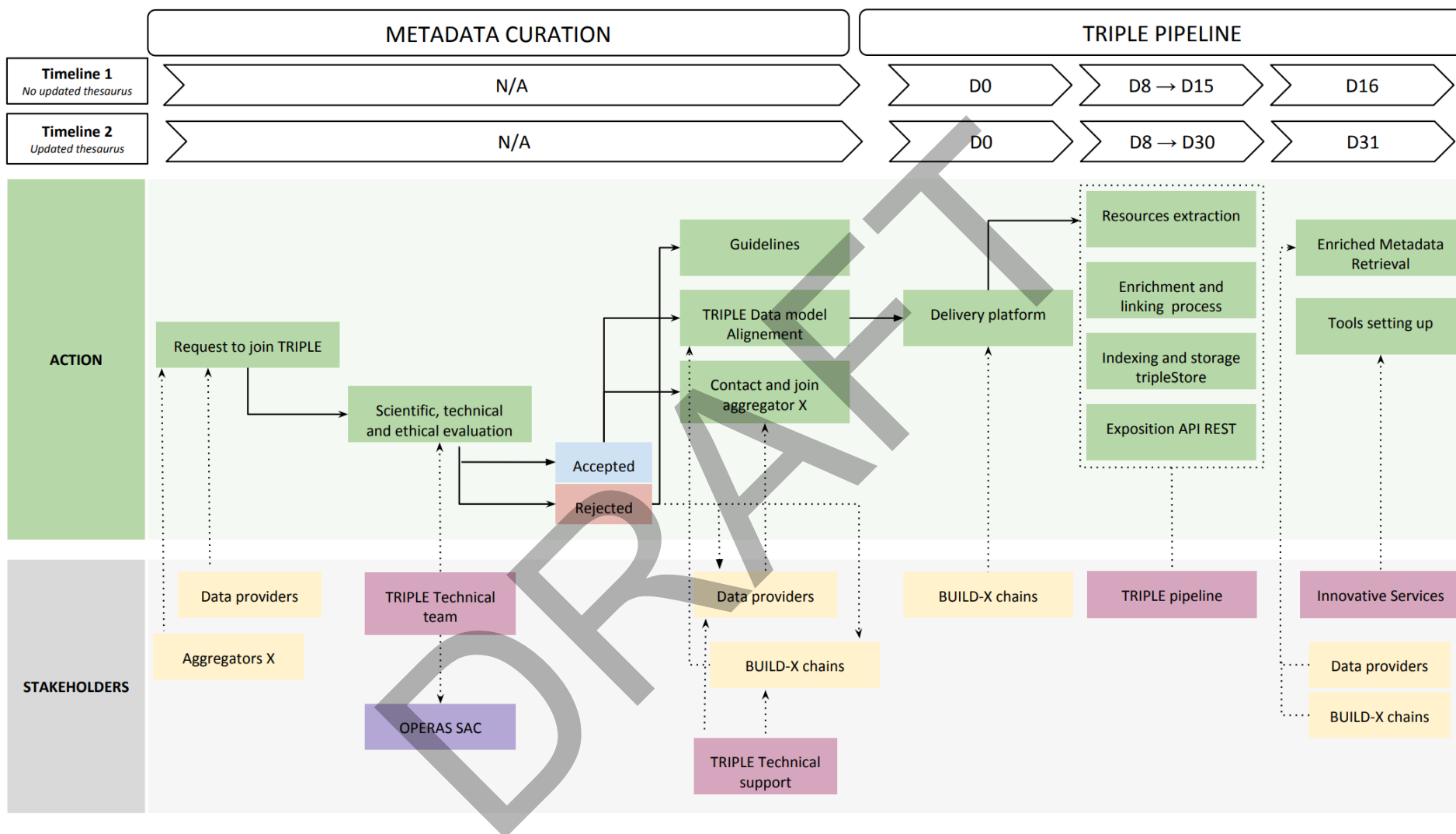


FIGURE 8. CHRONOLOGY TO ACQUIRE METADATA FROM THE DATA PROVIDERS UNTIL THEIR EXPOSITION TO USERS

7 | CONCLUSION

The Data Acquisition Plan sets out an ambitious blueprint for aggregating Social Science and Humanities data descriptions on a vast scale, in order to make many disparate data collections searchable and thus easily accessible to researchers via a single portal, which will constitute a part of EOSC. It provides a detailed approach in 2 phases to collect metadata in order to achieve this ambition. One of the challenges to be overcome is the fact that several metadata standards have to be mapped onto the common TRIPLE data model, which forms a key enabler for the TRIPLE platform. This '*lingua franca*' is vital for the running of the Innovative Services. The mapping task should be realized by aggregation platforms from their own data model. In many cases, the project (T2.2) will provide support to the data providers to help them remedy defects at source.

DRAFT

8 | GLOSSARY

Annotation

User annotation refers to personal online editing of any web content, such as: highlighting, inserting comments and notes.

Annotation for the TRIPLE enrichment process refers to adding structure to data (metadata or content) by creating links between terms (not concepts) and controlled vocabularies (taxonomies).

Aggregation platform

A platform which allows users to access publications, projects or user profiles. There are currently many platforms that cover national or international areas

API (Application Programming Interface) [2]

A language and message format used by an application program to communicate with the operating system or some other control program such as a database management system (DBMS) or communications protocol. APIs are implemented by writing function calls in the program, which provide the linkage to the required subroutine for execution.

BUILD processing chain (or BUILD chain)

A chain of software components developed by an operator to offer an aggregation platform.

Cardinality

Specification of the number of times that a metadata element can or must appear in a metadata description.

Data (or metadata) model [4]

An abstract model or representation of data for a particular domain, business enterprise, or field of study, independent of any specific software or information system. Usually expressed in terms of entities and relationships.

Controlled vocabularies [1]

A list of terms that have been enumerated explicitly, which may include relationships among them. Their purposes are to facilitate translations, promote consistency in the use of terms, indicate

semantic relationships among terms, label and browse the terms according to hierarchies for navigation systems, and enhance retrieval.

Crosswalk

A mapping of the elements, semantics, and syntax from one metadata schema to those of another. A crosswalk allows metadata created by one community to be used by another group that employs a different metadata standard.

Data element (or metadata element) [3]

A unit used by any given metadata standard to describe a resource. They correspond to fields in databases (e.g. author, title, rights) or properties in an RDF model (e.g. hasAuthor).

Data provider [4]

In Open Archives Initiative nomenclature, an organization that exposes metadata records in one or more repositories (specially configured servers) for harvesting by service providers.

Harvester [10]

A client application that issues protocols requests. A harvester is operated by a service provider as a means of collecting metadata from repositories.

Harvesting process

A mechanism that allows metadata to be collected from a remote catalog (or a remote database) and stored in a local space (server) for faster access. This harvest (or harvesting) is done regularly and automatically. To perform this harvest, the harvesting organization must use the same protocol techniques. To harvest in the open archives, the OAI PMH protocol is used.

Interoperability [4]

The ability of different information systems to work together, particularly in the correct interpretation of data semantics and functionality.

Linked Open Data (LOD) [2]

Refers to using the Web to automatically connect related data that previously were not linked in order to enhance discovery, retrieval and reuse and promote knowledge. To do so requires expressing metadata in a standardized and machine-readable manner.

Mapping

Action which translates elements and values from one schema to those of another. Mapping can lead to the creation of crosswalks that facilitate interoperability between different metadata schemas and serve as the base for metadata harvesting and exchange.

Metadata [5]

Data about data. Specifically, it is machine-readable data that describes content, context and structure of resources and their management over time. For this reason, all metadata is descriptive in nature, although other types may exist according to their function:

- Descriptive metadata: describes a resource for purposes of discovery and identification (author, title, creation date, language, etc.);
- Administrative metadata: provides information to help manage a resource. Includes rights metadata (licenses, copyright), technical metadata (software and hardware documentation) and preservation metadata (author of metadata, date and location, standards used, access privileges);
- Structural metadata: describes the relationships of parts of resources in relation to one another (e.g. the page number in a reading software).

Metadata record [6]

A set of metadata that describes a resource. It is comprised of syntax, structure and semantics:

- The syntax is a container for the structure and semantics, acting as a set of rules by which the contents should be interpreted by a computer (e.g. RDF, XML);
- The structure consists of the metadata scheme (e.g. Dublin Core), enabling humans to interpret the information within data elements (e.g. creator, title, subject);
- The semantics correspond to the content standards and define how the information within the data elements should be formatted (e.g. 01-05-1990 reads 1st May 1990 and not 5th January 1990).

Open Data [8]

Data is open if it can be freely accessed, used, modified and shared by anyone for any purpose - subject only, at most, to requirements to provide attribution and/or share-alike. Specifically, open data is defined by the Open Definition and requires that the data be A. Legally open: that is, available under an open (data) license that permits anyone freely to access, reuse and redistribute B. Technically open: that is, that the data be available for no more than the cost of reproduction and in machine-readable and bulk form.

Operator

An organization which manages an aggregation platform. He can be public, private or associative.

Persistent identifier (PID) [9]

A long-lasting reference to a digital resource.

Pipeline

Series of data processing steps. It comprises:

- Collecting or extracting raw data sets
- Data governance
- Data transformation (Standardization, Deduplication, Verification, Classification, Data sharing)

Protocol [4]

A specification—often a standard—that describes how computers communicate with each other (e.g., the TCP/IP suite of communication protocols or Open Archives Initiative Protocol for Metadata Harvesting [OAI-PMH]).

Repository [10]

A network accessible server that can process the protocols requests. A repository is managed by a data provider to expose metadata to harvesters.

Resource

Anything that can be described by metadata, regardless of its nature (physical or digital) or form (book, work of art, person).

Standard [7]

A formal document that establishes uniform criteria, methods, processes and practices. Metadata-related standards are created to guide the design, creation and implementation of data structures (Dublin Core, MARC21), data values (ISO 639-2), data contents (RDA, AACR2) and data exchange (RDFa, XML, JSON) in an efficient and consistent manner.

Semantic enrichment

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

Technical specification

A detailed description of technical requirements, usually with specific acceptance criteria, stated in terms suitable.

DRAFT

9 | REFERENCES

- [1] ANSI/NISO. (2006). Z39.87-2006 Data Dictionary – Technical Metadata for Digital Still Images.
- [2] DDI. [Online]. Available on: <https://ddialliance.org/resources/ddi-glossary>. [Accessed May, 2020].
- [3] Haynes, D. (2018). Metadata for information management and retrieval (2). London: Facet Publishing.
- [4] Introduction of metadata. [Online]. Available on: <https://www.getty.edu/publications/intrometadata/glossary/>. [Accessed June, 2020].
- [5] ISO 15489-1. (2016). Information and documentation — Records management. Geneva: ISO.
- [6] Lucas, R., Jackson, A., & Schneider, I. (2013). The metadata manual: a practical handbook. Oxford: Chandos Publishing.
- [7] NISO. (2017). Understanding metadata: what is metadata, and what is it for? Bethesda: NISO Press.
- [8] Open data handbook. [Online]. Available on: <http://opendatahandbook.org/glossary/en/terms/open-data/>. [Accessed May, 2020].
- [9] ORCID. [Online]. Available on: <https://support.orcid.org/hc/en-us/articles/360006971013-What-are-persistent-identifiers-PIDs->. [Accessed June, 2020].
- [10] The Open Archives Initiative Protocol for Metadata Harvesting. [Online]. Available on: <http://www.openarchives.org/OAI/openarchivesprotocol.html#DefinitionsConcepts>. [Accessed May, 2020].
- [11] Transforming Research through Innovative Practices for Linked interdisciplinary Exploration TRIPLE. The European discovery platform dedicated to SSH resources. Proposal number: 863420 (2018).