Triple

Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration

[12/2020]

Advancing Open Scholarship

D2.2 – DATA HARVESTING BEST PRACTICES DOCUMENT FOR DATA PROVIDERS

Version 1.0 – Final/PUBLIC

DATA HARVESTING BEST PRACTICES DOCUMENT FOR DATA PROVIDERS

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	WP2
WP Leader:	**IBL-PAN**
Deliverable identifier	
Contractual Delivery Date: 31/12/2020	**Actual Delivery Date:** 21/12/2020
Nature: Report	**Version: 1.0 Final**
Dissemination level	**PU**

Revision History

Version	Created/Modifier	Comments
0.1	Melanie Bunel, Laurent Capelli, Luca De Santis, Suzanne Dumouchel, Haris Geogriadis, Arnaud Gingold, Twan Goosen, Alexander König, Maciej Maryl, Andre Pacheco, Panayiota Polydoratou, Jan Rohdan, Jadranka Stojanovski, Tomasz Umerle	Internal review
0.2	Maxime Bouillard, Arnaud Roi	External review

| 1.0 | Melanie Bunel, Laurent Capelli, Luca De Santis, Suzanne Dumouchel, Haris Geogriadis, Arnaud Gingold, Twan Goosen, Alexander König, Maciej Maryl, Andre Pacheco, Panayiota Polydoratou, Jan Rohdan, Jadranka Stojanovski, Tomasz Umerle | Final version |

Table of Contents

List of Figures

List of Tables

Acronyms

DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
DCTerms	DCMI Metadata Terms
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OPERAS	European Research Infrastructure for the development of open scholarly communication in the social sciences and humanities
ORCID	Open Researcher and Contributor ID
RDFa	Resource Description Framework in Attributes
SSH	Social Sciences and Humanities
TRIPLE	Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration

Data Harvesting Best Practices Document for Data Providers

Publishable Summary

The TRIPLE project intends to address the challenges of the fragmented and heterogeneous landscape of Social Sciences and Humanities, hereafter SSH, research outputs' dissemination by providing a single yet flexible architecture able to collect data from many sources. Enabling the data ingestion from different data models, the architecture relies itself on an ontology which ensures a better discoverability for SSH scholarly resources and their connection with projects and profiles.

The proposed architecture aims at enabling both a broader collection of data, and a uniform quality level between these through the use of a single data model. In order to do so, the architecture distinguishes two main roles, the TRIPLE's Aggregator, and the TRIPLE's Provider. The Aggregator is the contact point between the Provider and TRIPLE, and takes charge of the mapping between its own data and the TRIPLE data model. This architecture thus ensures that the discovery platform will provide a large coverage of existing SSH scholarly resources, with Aggregators providing, where necessary, the useful support to its own Providers.

This document is intended as a beta version of the public document that will provide guidelines to the TRIPLE's data providers. The final public guidelines will be a mix of this document and of the Data acquisition plan. Although the core of the guidelines is not expected to change significantly over time, it should be noted that as the GOTRIPLE platform evolves, adjustments and additions might be required to facilitate the data provision system described below.

I. INTRODUCTION

The TRIPLE project intends to address the challenges of the fragmented and heterogeneous landscape of Social Sciences and Humanities, hereafter SSH, research outputs' dissemination by providing a single yet flexible architecture able to collect data from many sources. Enabling the data ingestion from different data models, the architecture relies itself on an ontology which ensures a better discoverability for SSH scholarly resources and their connection with projects and profiles.

Throughout this document, when we refer to data (scholarly resources, projects and profiles), we refer to the corresponding metadata that will be, when possible, openly accessible under CC0 waiver. Where applicable, the metadata will include a link to facilitate access to the actual data (e.g., the full text of the publications).

As the deliverable offers guidance to the TRIPLE data providers, it is focused on the data which directly implies their action, e.g., the research outputs, hereafter called scholarly resources. For the sake of completeness, the automated collection of data regarding profiles and projects will be also quickly mentioned.

This document should be read alongside with the TRIPLE Data Acquisition Plan[1]. The acquisition plan details the TRIPLE architecture.

The proposed architecture aims at enabling both a broader collection of data, and a uniform quality level between these through the use of a single data model. In order to do so, the architecture distinguishes two main roles, the TRIPLE's Aggregator, and the TRIPLE's Provider. The Aggregator is the contact point between the Provider and TRIPLE, and takes charge of the mapping between its own data and the TRIPLE data model. This architecture thus ensures that the discovery platform will provide a large coverage of existing SSH scholarly resources, with Aggregators providing, where necessary, the useful support to its own Providers.

The distinction between Providers and Aggregators determines the structure of this document. A section (section III) is dedicated to the Providers guidelines. The guidelines describe the actions that the Providers should take to ensure the visibility of their content according to general best practices for data management, such as the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The requirements are also more specifically related to TRIPLE's innovative services (Recommender System, Visualization tool, Annotation tool, Trust Building System).

The primary Aggregator for TRIPLE is the platform ISIDORE, managed by Huma-Num, coordinator of the project. ISIDORE is the basis for the development of the TRIPLE architecture and will offer specific support to its Providers. The section dedicated to the Providers guidelines is therefore focused on the Providers entering TRIPLE through the Aggregator ISIDORE. It is however intended

---

[1] https://www.gotriple.eu/?page_id=42.

Data Harvesting Best Practices Document for Data Providers

that these guidelines will give useful information also to the other Aggregators that will collaborate with TRIPLE.

Another section (section IV) provides guidance on harvesting by other Aggregators. At this stage, this section presents the preliminary results of a survey about potential TRIPLE Aggregators. This section will help the TRIPLE consortium to assess the specific actions to conduct in order to explore collaboration with each of these Aggregators. The section also gives useful information to data managers either to select the best route to enter TRIPLE, or to assess the possibility of becoming a candidate-Aggregator for TRIPLE themselves.

This document is intended as a beta version of the public document that will provide guidelines to the TRIPLE's data providers. The final public guidelines will be a mix of this document and of the Data acquisition plan. Although the core of the guidelines is not expected to change significantly over time, it should be noted that as the GOTRIPLE platform evolves, adjustments and additions might be required to facilitate the data provision system described below.

II. SERVICES AND DEFINITIONS

The TRIPLE's services and definitions are here provided in the perspective of Providers' and Aggregators' data acquisition. TRIPLE, as a whole, relies on an entry point (the Delivery platform), an enrichment pipeline (TRIPLE Core), and an end-user interface (the GOTRIPLE discovery platform).

The main steps and roles are schematically represented in the diagram below; they will be further described in this section.
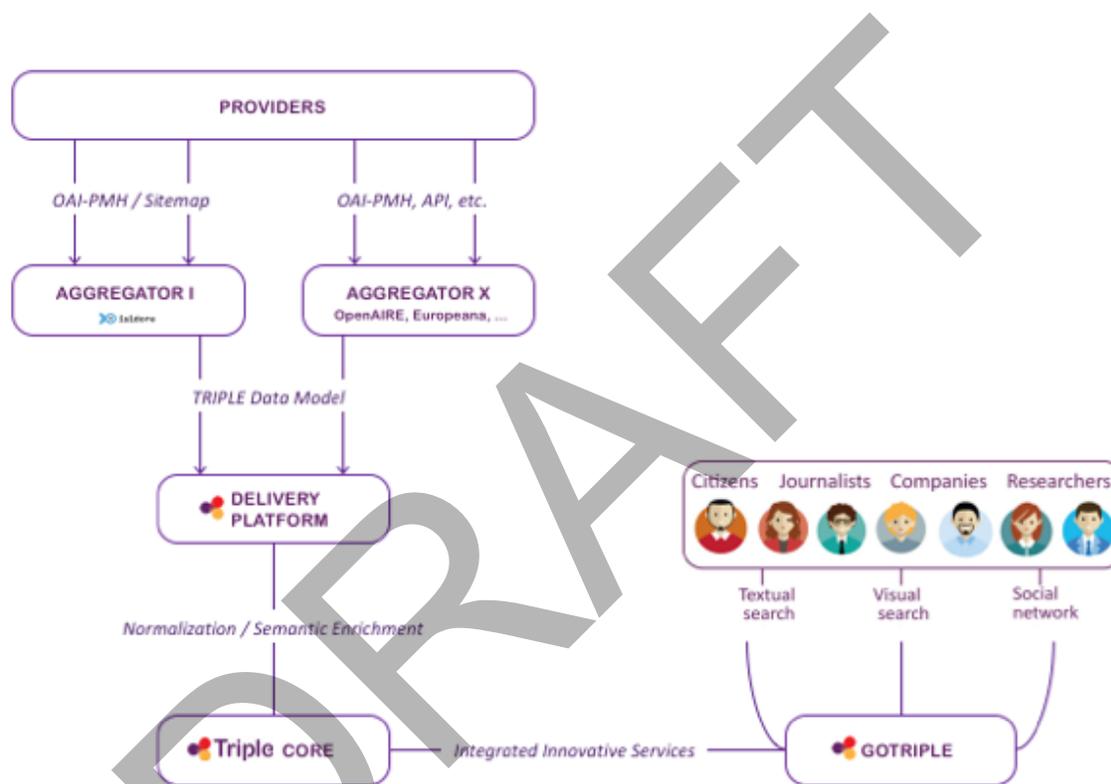
FIGURE 1: SCHEMATIC VIEW OF THE DATA ACQUISITION PROCESS IN TRIPLE

A. Discovery content and features

1. Scholarly resources

The main focus of this deliverable are the scholarly resources in the fields of social sciences and humanities. Data about projects and profiles (see section A.2 below) will be collected automatically and therefore do not involve the data providers for which these guidelines are designed.

Data Harvesting Best Practices Document for Data Providers

Such scholarly resources include but are not limited to different kinds of written outputs (books, scholarly articles, archival resources, etc.) and different resource data types (for example audio visual resources, datasets, data papers, images, etc.). These scholarly resources will be brought into the TRIPLE environment by various stakeholders in the SSH ecosystem (see section C), and made available to the public through the GOTRIPLE platform.

The aim of these guidelines is to help the providers of data to deliver the best quality of metadata thus increasing their resources' discoverability. The TRIPLE data model ensures that the collected metadata can be adequately enriched, indexed, and processed by our innovative services. These objectives determine the distinct priority levels of each information that the data providers will have to make available for TRIPLE.

2. Research projects and researchers' profiles

Discovery will be further facilitated by including metadata associated with research projects and researchers' profiles[2]. The TRIPLE project will thus address the lack of existing tools and aggregators to discover these data types. The TRIPLE ontology will help to increase the metadata quality for these types of data.

In this context, various options are envisioned. The research projects will be incorporated, wherever applicable, via the following paths:

Research projects:

- European platforms, like OpenAIRE[3],
- National funding agencies databases;
- Scientific projects, funded or not, led by SSH researchers.
- The crowdfunding tool.

As far as the latter point is concerned, in the TRIPLE project, as planned in its Description of Work, an integration with an existing crowdfunding service will be pursued. Crowdfunding can represent in fact a promising way to support research on a wide variety of SSH-related topics, allowing a quick and effective solution for a project to receive funds and therefore encouraging the analysis even of "smaller" but possibly quite "urgent" issues, that can have a wide impact on society. While work on this topic has just started in T5.1, initially focusing on the analysis of legal and financial issues, the final goal is to design and implement a full integration of the GOTRIPLE platform with an existing crowdfunding service, still to be identified. All the descriptions of its projects related to the SSH domain will be automatically imported in GOTRIPLE and made available through its discovery functionality.

Regarding the researchers' profiles, here are the potential sources currently considered:

---

[2] More information is available in D2.1 "Data acquisition plan".
[3] https://www.openaire.eu/

- Aggregation platforms which collect profiles or allow to create profiles on their platforms (for example, scanR[4]);
- the ORCID registry,
- and GOTRIPLE platform's own users who will be invited to register their profile should they wish to do so.

With the exception of the cases in which the information can be collected directly from the provider's metadata records, the processes of gathering metadata about profiles and projects are distinct and out of the specific scope of these guidelines.

B. Sources: Providers and Aggregators

In order to facilitate and expand data harvesting, the TRIPLE team has designed a specific architecture allowing to gather more data from more sources. In this context, TRIPLE architecture distinguishes two distinct platforms: the Delivery platform, and the End-user platform. The Delivery platform is the internal place where the data is received and stored before entering the TRIPLE enrichment pipeline. The End-user platform, GOTRIPLE, is the public interface where the enriched data is made available[5]. The TRIPLE core entails all the enrichment processes taking place between the Delivery platform and the End-user platform. The Delivery platform, the TRIPLE core, and GOTRIPLE constitute the TRIPLE system.

The Delivery platform is the endpoint for distinct data acquisition processes set with specific actors, here called Aggregators. Each Aggregator delivers metadata harvested from its own sources, here called the Providers. The definition of Aggregators and Providers is primarily functional rather than technical: they relate to the specific interaction, direct or indirect, with the Delivery platform. Only Aggregators interact with the Delivery platform. The Providers interact only with the Aggregators.

A TRIPLE's Aggregator is an organization that collects, manages, and disseminates the metadata of the scholarly resources made available by various Providers. The Aggregator operates as a standardization body of heterogeneous metadata, either by defining its own requirements, or by relying on existing standards for harvesting and dissemination. In that sense, a federation of Providers is not considered to be an Aggregator if the standardization process is not centralized. An Aggregator may set up specific processes to further enrich the metadata collected; in that prospect, it may necessitate access to the original data to conduct specific enrichment processes. According to the TRIPLE's criteria, an Aggregator operates in the field of public research and provides openly accessible data. An aggregator can assume a variety of types regarding: the perimeter (regional, national, European), the scope (geographic, thematic), the mission (official mandate, service provision), and the status (legal entity, project, association). TRIPLE Aggregators[6]

---

[4] https://scanr.enseignementsup-recherche.gouv.fr/recherche/persons
[5] Although it doesn't have direct implication on the data provision or collection, it should be noted that the enriched data will also be available through a SPARQL endpoint.
[6] The current list of TRIPLE potential Aggregators is available in section IV.

Data Harvesting Best Practices Document for Data Providers

are validated according to the TRIPLE's admission criteria, after consultation of the TRIPLE Advisory Board in specific cases (see section V).

A TRIPLE's Provider is an organization that manages, collects, and disseminates scholarly resources. It operates as the manager of one or various data repositories, archives, or publishing platforms. 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. A Provider doesn't have a specific legal status but provides the scholarly resources according to the open access requirements conforming to the TRIPLE's policy. Conventionally, a Provider provides its data to an Aggregator, generally through the OAI-PMH[7] protocol; an Aggregator sets up a data acquisition process with TRIPLE. From the end-user perspective, this data acquisition process is seamless, and TRIPLE will manage, enrich, and expose all the collected data in a uniform way, thanks to the TRIPLE data model common for all scholarly resources (see section III).

This document is structured according to the TRIPLE's distinction between Providers and Aggregators. A section is therefore dedicated to the Providers guidelines. ISIDORE is the primary Aggregator for TRIPLE and these guidelines are directed towards the ISIDORE providers. A second section provides guidance on harvesting by other Aggregators. The document also offers useful information to data managers to assess the possibility for their repository to become a candidate-Aggregator for TRIPLE. TRIPLE partners themselves have the possibility, depending on their technical service and readiness, to become a TRIPLE's Aggregator.

From the data managers' perspective, the architecture allows therefore 3 options:

1. being harvested by the TRIPLE Aggregator ISIDORE;
2. being harvested by one of the other selected Aggregators,
3. becoming a TRIPLE Aggregator.

As stated in the introduction, these guidelines provide details on point 1) in section III. They also provide guidance on point 2) in section IV. The whole document offers therefore useful information to data managers, who would like to take steps towards point 3).

C. Scope: TRIPLE, OPERAS and beyond

The TRIPLE architecture allows to broaden the number of sources: although the sources of TRIPLE will primarily come from the project's partners and the members of OPERAS, it is also intended to provide a discovery solution to independent Providers. More broadly, the TRIPLE architecture allows as well to gather metadata from Aggregators external to the project.

The first source and Aggregator of TRIPLE is the ISIDORE platform, as it is the ground on which TRIPLE is built. Additional sources come from or are identified by the TRIPLE project's partners: CESSDA, CLARIN, EKT, IBL PAN, MWS, University of Coimbra, University of Zadar. After checking

---

[7] The "Open Archive Initiative - Protocol for Metadata Harvesting (OAI-PMH)" is an open protocol to harvest metadata from repositories conforming to OAI requirements.

whether they can be an Aggregator for TRIPLE, these partners will assess the best route for the harvesting of their data by the TRIPLE's Aggregators, starting from ISIDORE.

Other sources will be identified within the OPERAS network (see Annex I). OPERAS comprises to date 42 partners representing a wide range of actors in the SSH ecosystem, such as researchers, publishers, libraries and archives. Scholarly resources produced, curated and hosted by the OPERAS partners will be incorporated in the TRIPLE system according to the options detailed in the present guidelines.

Crucial to the TRIPLE architecture, additional sources will come from Aggregators external to the project. Based on the definitions provided above, the TRIPLE team will identify potential Aggregators within its scope: open scholarly resources in the SSH fields. One of the objectives of this deliverable is to provide preliminary information on these potential Aggregators. The current list is detailed below, under section IV.

Potential Providers include any actors in the SSH ecosystem who wish to make their metadata for scholarly resources visible and accessible. Particular care and support will be provided to all actors who make their scholarly resources available via the "diamond model"[8] access routes (e.g. in diamond journals and platforms). These actors should be primarily oriented towards the Aggregator ISIDORE.

D. Admission criteria for scholarly resources data

- Content: Scholarly resources in SSH, openly accessible, as defined in section II A.
- Resources types: the current list of data types that will be referenced in GOTRIPLE conforms to the list of resource types of ISIDORE, which is based on the related COAR controlled vocabulary.[9]
- Licensing: Metadata of such resources should be made available, as far as possible, under CC0 waiver, or an open license allowing for its full reuse by TRIPLE.
- Technical compliance: Technical requirements compliance as these are described in section III A.
- Metadata compliance: Metadata compliance based on requirements described in section III B.
- Aggregator for TRIPLE: Operates in the field of public research and provides openly accessible data.

---

[8] https://www.coalition-s.org/exploring-collaborative-non-commercial-publishing-models-for-open-access/
[9] http://vocabularies.coar-repositories.org/documentation/resource_types/. The list of resources types will be updated in the final version of the guidelines.

Data Harvesting Best Practices Document for Data Providers

III. PROVIDERS: BEST PRACTICES FOR HARVESTING BY ISIDORE

Providers who cannot become Aggregators, or find a suitable Aggregator in their environment, can have their content exposed in GOTRIPLE through the ISIDORE platform. The following sections describe the guidelines for Providers that will facilitate their harvesting by ISIDORE.

A. Harvesting protocol

ISIDORE harvests scholarly resources metadata accessible on the web and in open and interoperable standards (Dublin Core). ISIDORE enriches metadata in 3 languages: French, English and Spanish.

This harvesting process is done by ISIDORE regularly and launched manually by the developers, at least once a month in order to get the latest updates. ISIDORE 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.

1. OAI-PMH protocol

The OAI-PMH protocol was developed in 1999 as part of the Open Archives Initiative. 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, ListIdentifiers, ListMetadataFormats, ListRecords and ListSets. The OAI-PMH protocol implies 2 stakeholders: 1) The data provider, that creates structural metadata and exposes them in a repository for harvesting and 2) The service provider, that harvests 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 Core 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 ISIDORE 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 identifier. The delete option of the OAI protocol is also considered by the harvesting process. To update 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 ISIDORE 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 configured to accept or reject some OAI sets.

2.   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 1).

```
<?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 2: EXAMPLE OF A SITEMAPS XML FILE

In the context of the ISIDORE 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.

Therefore, 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 ISIDORE harvester which would be limited to specific pages containing relevant scientific content for TRIPLE. 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 "Lastmod" field (for "last modification") 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 ISIDORE. ISIDORE retrieves RDFa metadata in the Dublin Core and DCMI Metadata Terms standards.

Data Harvesting Best Practices Document for Data Providers

B. Metadata compliance

In order to ensure high semantic expressivity and address flexibility needs, the TRIPLE data model is based on the Schema.org ontology. The semantics of Schema.org 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 is recommended to the TRIPLE's Aggregators that will have to deliver their metadata packets according to this model. It applies therefore to the Providers providing their data through the Aggregator ISIDORE.

The TRIPLE data model allows for flexibility by enabling the collection of heterogeneous data: as such, there are no requirements regarding the metadata standards used by the Providers, as long as their metadata can be collected through the 2 protocols described above. Furthermore, as specified before, the TRIPLE data model accounts for different types of scholarly resources, which can all be described with a generic model, only specifying the type.

In this prospect, the number of mandatory elements are limited and focused on minimal indexing requirements: *Title, Identifier, Creator (Author)*.

Other elements are marked as recommended because they are closely related to the GOTRIPLE specificities. Firstly, certain elements are necessary for some TRIPLE Innovative services, namely the visualisation tools: *Language, Keywords, Description (abstract)*. Other elements relate to the focus on open access that defines TRIPLE's mission: *Landing page of the document*, *Link to the full text* (when applicable), *License*. Finally, since the GOTRIPLE platform is intended to link the discovery of scholarly resources and their related projects, the following element is also recommended: *Funding reference* (including relation with a research project).
All other metadata elements are optional but ensure the richness of metadata according to the FAIR principles. More specifically, such elements enable Providers to have their metadata aligned with the major existing guidelines in the research environment, e.g., OpenAIRE or Datacite.
Good quality metadata ensures higher visibility for the data and it is therefore recommended to ensure completeness of records, conformance to standards, accuracy and consistency, etc. as much as possible.

The table below summarizes the different metadata elements which can be used to describe scholarly resources, ordered according to their level of priority for TRIPLE. The table provides a short description of the metadata element and its expression in the TRIPLE ontology. Where applicable, the mapping with the current data model from ISIDORE is also provided in the last column.

TABLE 1: TRIPLE METADATA ELEMENTS BY LEVELS OF PRIORITY

PRIORITY	ELEMENT DESCRIPTION	TRIPLE ONTOLOGY	ISIDORE mapping
Mandatory	Title of the resource	schema:headline	dcterms:title
Mandatory	Author of the resource	schema:author	dcterms:creator
Mandatory	Identification of the resource through a string or a number conforming to a formal identification system	schema:identifier	dcterms:identifier
Recommended	Language of the content of the resource	schema:inLanguage	x
Recommended	Keywords or tags used to describe this content	schema:keywords	dc:subject
Recommended	Textual description of the content	schema:abstract	dc:description
Recommended	Link to the landing page of the resource	schema:mainEntityOfPage	ore:aggregates
Recommended	Link to the full resource (URL) with AccessRights (not necessarily Open Access)	schema:url	ore:aggregates
Recommended	A license document that applies to this content, typically indicated by URL.	schema:license	x
Recommended	A reference to another publication or dataset	schema:citation	x
Recommended	An existing relation with a research project	schema:about	x
Recommended	Date of publication/creation	schema:datePublished	dc:date
Optional	Type of the resource	schema:additionalType	dc:type
Optional	Date of publication/creation	schema:datePublished	dc:date
Optional	Entity responsible for making the resource available	schema:publisher	dcterms:publisher

Data Harvesting Best Practices Document for Data Providers

Optional	Any other information about the resource	schema:mentions	dc:relation
Optional	A reference to a resource from which the present resource is derived.	schema:isBasedOnURL	dcterms:isPartOf
Optional	Coverage information including 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	dc:coverage dcterms:spatial

The Providers of ISIDORE will thus be encouraged and supported to adopt these guidelines ensuring the highest level of discoverability within the GOTRIPLE platform and in other important Aggregators.

The future version of the guidelines will provide more details of the requirements for each metadata element.

C. Licensing compliance

In order to facilitate the dissemination and reuse, GOTRIPLE will apply the CC0 waiver to the metadata it will expose. Such policy can indeed ensure that reuse will be possible where copyright might be included, and the CC0 waiver will simply have no effect when copyright will not be included. GOTRIPLE will thus encourage Providers to have compliant policies.

D. Quality control

A monitoring tool for the ISIDORE's Providers will be developed to help them verify the correct harvesting of their metadata and its appropriate dissemination. The tool will be developed with a coordinated effort of Task 2.6 (Harvesting Management) and Task 4.3 (Procurement and configuration of the back-office system powering the TRIPLE system).

E. Terms of use

Terms of use will be communicated to the ISIDORE's Providers, informing and requesting their agreement upon: ISIDORE's harvesting and enrichment processes, GOTRIPLE metadata licensing, quality assurance, termination of the harvesting.

IV. AGGREGATORS: PREPARATORY STUDY

A. List of aggregators: first list

Among the aims of this deliverable were to identify data acquisition actors outside TRIPLE that are already aggregating content of SSH scholarly resources. The criteria for doing so were: coverage of SSH scholarly resources, open access metadata, preferably available under CC0 waiver, and any existing coverage of TRIPLE partners content already.

The following were indicated by TRIPLE partners as those data acquisition actors that are either already collaborating for making their data visible or should be further explored. The list will be further extended during the course of the project, thanks to the partners' input about potential aggregators in their environment (geographic or thematic).

- BASE[10]: BASE is a search engine (search index based on Solr/Lucene) for academic web resources. It provides more than 150 million documents from more than 7,000 sources. BASE is operated by Bielefeld University Library. It indexes the metadata of all kinds of academically relevant resources – journals, institutional repositories, digital collections etc. – which provide an OAI interface and use OAI-PMH for providing their content.

- DOAB[11]: The Directory of Open Access Books is a joint service of OAPEN, OpenEdition, CNRS and Aix-Marseille Université. The primary aim of DOAB is to increase discoverability of Open Access books. The directory is open to all publishers who publish academic, peer reviewed books in Open Access and meet academic standards.

- DOAJ[12]: The Directory of Open Access Journals is a community-curated online directory that indexes and provides access to high quality, open access, peer-reviewed journals.

- Europeana[13]: The Europeana platform facilitates submission of data via a) domain/thematic aggregators: Domain & thematic aggregators define their scope by cultural sector (such as museum, archive or library) or by topic and theme (such as fashion). They work with contributors based in different European countries and b) national/regional aggregators: National & regional aggregators define their scope by specific country or region. They work with contributors situated within that country or region.

---

[10] https://www.base-search.net/
[11] https://www.doabooks.org/
[12] https://doaj.org/
[13] https://pro.europeana.eu/page/aggregators#domain-thematic-aggregators

Data Harvesting Best Practices Document for Data Providers

- MOSA[14]: In order to highlight the Open Access research published by Belgian and Luxembourgish scholars, University of Liège Library has launched Mosa, a project to harvest content from existing open access repositories in both countries. Mosa aims to act as an aggregator of institutional and federal publication platforms.

- NARCIS[15]: NARCIS.nl is the national portal for anyone looking for information about researchers and their work in The Netherlands. Besides researchers and policy makers, NARCIS is also used by students, journalists and education, government and business professionals. NARCIS is already harvested by OpenAIRE.

- OpenAIRE[16]: The OpenAIRE infrastructure enables content submission via various routes which are described in the aggregation and content provision workflows. One of them is the "Aggregator services" which are defined as information systems that collect descriptive metadata about publications or datasets from multiple sources in order to enable cross-data source discovery of given research products. Examples are DataCite, BASE, DOAJ.

B. Metadata specifications

- BASE[17]: BASE indexes metadata which provide an OAI interface and use OAI-PMH for providing their contents. Content providers should abide by the "Golden Rules for Repository Managers" as established by BASE, in order to ensure their indexing.

- DOAB[18]: DOAB supports the OAI protocol for metadata harvesting (OAI-PMH). Service providers and libraries can use the protocol to harvest the metadata of the records from DOAB for inclusion in their collections and catalogues. All books in DOAB are available in Dublin Core format, oai_dc format, and in MARC XML format.

- DOAJ[19]: DOAJ currently supports only oai_dc. However, detailed documentation (API, XML, OAI-PMH, widgets, data dumps,OpenURL, etc.) can be found at the DOAJ documentation page.

- Europeana[20]: Metadata submitted to the Europeana platform must conform to the Europeana Data Model and the content must align with one of the four content tiers and

---

[14] https://mosa-research.be/
[15] https://dans.knaw.nl/en/about/services/narcis
[16] https://www.openaire.eu/about
[17] https://www.base-search.net/about/en/faq_oai.php#chap04
[18] https://www.doabooks.org/doab?func=about&uiLanguage=en#metadata
[19] https://doaj.org/
[20] https://pro.europeana.eu/post/publishing-framework

three metadata tiers for quality outlined in the document [Europeana Publishing Framework: Metadata](#).

- [MOSA](#)[21]: Data can be submitted to MOSA via OAI-PMH protocol, FTP or manual data delivery. Deciding on the best option should be discussed with the MOSA team. In terms of supported formats those include Dublin Core and XML.

- [NARCIS](#)[22]: In order to facilitate data exchange NARCIS places agreements between data providers and service providers. All of the technical requirements can be found at http://harvester.dans.knaw.nl/. They relate to the use of:
  - The OAI-PMH exchange protocol;
  - The URN:NBN persistent identifier;
  - Semantics within the MODS/DIDL format;
  - The MODS metadata format;
  - The DIDL metadata format.

  Metadata may also be submitted in Dublin Core format, but in that case the URN:NBN persistent identifier cannot be used and author identifiers such as DAI, ISNI or ORCID are not supported.

- [OpenAIRE](#)[23]: OpenAIRE provides guidelines for different data acquisition actors (e.g. repositories, data archives, research information systems, software repositories, etc.). OpenAIRE guidelines are based on a data model that uses elements from DublinCore and Datacite, together with elements specific to OpenAIRE. The OpenAIRE guidelines are compliant with EC recommendations for Open Access. The guidelines are specific to data types: there are distinct guidelines for literature repositories and data archives. The guidelines can also differ from one release to another (latest version for literature repositories is V4.0).

C. Licensing specifications

- **BASE:** No public information on this topic.

- [DOAB](#)[24]: All metadata feeds are available under a CC0 1.0 license.

---

[21] https://explore.lib.uliege.be/discovery/static-file/help?vid=32ULG_INST:MOSA
[22] https://www.narcis.nl/inclusion/Language/en
[23] https://guidelines.openaire.eu/en/latest/index.html
[24] https://www.doabooks.org/doab?func=about&uiLanguage=en#metadata

Data Harvesting Best Practices Document for Data Providers

- [DOAJ](#)[25]: The data in DOAJ is licensed to third parties under the Creative Commons Attribution-ShareAlike License (CC BY-SA).

- [Europeana](#)[26]: The Europeana Licensing Framework aims to standardise and harmonise rights. It comprises four elements: the **data exchange agreement** (*specifies how metadata and previews provided by data providers can be used by Europeana and third parties)*, the **CC0 waiver** (*The CC0 waiver is a legal tool that has been developed by Creative Commons for making data available without restrictions on re-use. The DEA establishes that Europeana publishes metadata it receives from its data providers under the terms of the Creative Commons Zero Universal Public Domain Dedication)*, the **Europeana terms for user contributions** (*These terms have been developed so that Europeana can use content provided by its users and be able to integrate with other Europeana held content and data*) and the edm:rights field for the **EDM model** (*Europeana uses the contents of the edm: rights to tell users under which terms they can use the previews and digital objects that they find by Europeana).*

- [MOSA](#)[27]: No public information on this topic.

- [NARCIS](#)[28]: Harvesting the information in NARCIS is permitted after registration. Re-use of information in sections "Persons" and "Organisations" is restricted.

- [OpenAIRE](#)[29]: OpenAIRE describes the content acquisition policy, the terms of agreements with content providers and the acceptable use policy can be found [here](#). Enriched metadata can be available under the CC-BY v.4 licence.

D. Quality control

Quality control will be closely interlinked with tasks 2.6 and 4.3 that will provide some analytics through the creation of a monitoring tool at two levels:
- A monitoring tool in the Aggregator ISIDORE for the data providers to have information and statistics about the harvesting of their data;
- A monitoring tool in the GOTRIPLE platform for the Aggregators to have information and statistics about the data they deliver on the delivery platform. More information on this on-going work will be provided as technical development evolves. The objective is to

---

[25] https://doaj.org/docs/faq/#restrictionsA
[26] https://pro.europeana.eu/page/europeana-licensing-framework
[27] https://explore.lib.uliege.be/discovery/static-file/about?vid=32ULG_INST:MOSA
[28] https://www.narcis.nl/terms/Language/en
[29] https://www.openaire.eu/policies

create an interactive communication tool between the different stakeholders in order to improve the quality of metadata.

E.   BUILD chain set-up

The TRIPLE pipeline will not directly harvest metadata but it will exploit metadata provided by processing chains, called BUILD-X chains, managed by Aggregators which already harvest metadata. The BUILD-X chains will deposit packets of metadata on the Delivery platform at regular periods to be defined in negotiation with the selected Aggregators. 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". The TRIPLE data model will be a recommendation for the Aggregators (such as ISIDORE, OpenAIRE, NARCIS, etc.) which will have to deliver their metadata packets according to the model based on a subset of the schema.org ontology.

ISIDORE is called the BUILD-I chain and will be the first implementation of this architecture. Specific push-pull processes and scenarios for additional BUILD chains will be determined together with the concerned Aggregators.

F.   Location of data: a first cartography of the TRIPLE partners' content

As mentioned in section IV, the criteria we applied for identifying Aggregators external to TRIPLE were: the coverage of SSH scholarly resources, open access metadata, as far as possible available under CC0 waiver, and any existing coverage of TRIPLE partners' content already.

The following tables establish a cartography of the TRIPLE partner's current or potential data location among identified Aggregators. The table 2 shows where the TRIPLE partners content is currently available. The table 3 shows the potential different entry points into the TRIPLE system. This list of identified Aggregators and potential routes for Providers will serve as a basis for the establishment of direct collaborations and agreements with Aggregators external to TRIPLE in 2021.

Specifically, at the moment, only the OpenEdition content is available via the ISIDORE Aggregator. EKT, MWS, and University of Zadar content is accessible via OpenAIRE, and EKT cultural heritage content is accessible via Europeana. EKT, MWS, and OpenEdition content is also made accessible via BASE.

Data Harvesting Best Practices Document for Data Providers

TABLE 2: CURRENT AVAILABILITY OF TRIPLE PARTNERS' CONTENT

TRIPLE partners	Aggregators through which TRIPLE partners' content is currently available							
	ISIDORE	OpenAIRE	Europeana	NARCIS	MOSA	BASE	DOAJ	DOAB
CESSDA								
CLARIN								
COIMBRA							x	x
EKT		x	x			x		
IBL-PAN								
MWS		x				x		
OpenEdition	x	x				x	x	x
University of Zadar		x						

In the following table, we are presenting the options for all TRIPLE partners' content to enter the TRIPLE system based on its current architecture. Irrespective of whether a partner's content could enter the TRIPLE system via external Aggregators (e.g., OpenAIRE, Europeana, etc.), there is always the option to make their content visible via the Aggregator ISIDORE.

TABLE 3: CHANNELS THROUGH WHICH TRIPLE PARTNERS' CONTENT COULD ENTER GOTRIPLE IN THE FUTURE

TRIPLE partners	Aggregators through which TRIPLE partners' content could enter TRIPLE							
	ISIDORE	OpenAIRE	Europeana	NARCIS	MOSA	BASE	DOAJ	DOAB
CESSDA	x							
CLARIN	x							
COIMBRA	x						x	x
EKT	x	x	x			x		
IBL-PAN	x							
MWS	x	x				x		
OpenEdition	x	x				x	x	x
University of Zadar	x	x						

This preliminary survey allows to identify various challenges that should be addressed in future discussion with Aggregators:

- Content related challenges:
    - How to identify and isolate scholarly resources conforming to TRIPLE definition, openly accessible, and specific to SSH across different Aggregators;
    - How to address the difference of data types across different aggregators (e.g., literature and datasets distinction in OpenAIRE vs scholarly resources in TRIPLE).
- Metadata related challenges:
    - Level of data models' coincidence between the various Aggregators once mapped to the TRIPLE data model (e.g., the recommended fields might have very different coverage in the final TRIPLE database).
- Licensing related challenges:

o How to conform to licensing options across different aggregators (e.g., different licensing policies).

DRAFT

Data Harvesting Best Practices Document for Data Providers

V.   TRIPLE ADVISORY BOARD

In order to build an SSH discovery platform that conforms to scientific high-quality standards and enhance links with society as well, the TRIPLE consortium intends to rely on an external pool of experts. They will help the TRIPLE partners in specific questions which influence technical decisions.

This pool of experts is called "TRIPLE Advisory Board" and it will be consulted on concrete and pragmatic questions with the goal to build a platform for researchers and other stakeholders that will better meet their needs.

The primary body of consultation within the TRIPLE's governance model is the Technical Board (TB) involving all the partners taking part in the technical building of the TRIPLE system. Issues raised about technical implementations and policy options are first discussed and addressed in the TB. The TRIPLE Advisory Board (TAB) will be solicited upon request only, with specific questions prepared by the TB.

In the context of the data provision considered in this document, the TAB will be consulted on the addition of new Aggregators that will not fully match the expected criteria. The TAB may also give advice on the licensing policy adopted and the challenges it may face.

VI. ANNEXES

A. Annex I: List of OPERAS partners

Country	Name of partner	Acronym	URL
UK	Joint Information Systems Committee	JISC	https://www.jisc.ac.uk/
DE	Max Weber Stiftung	MWS	https://www.maxweberstiftung.de/startseite.html
GR	National Documentation Centre/National Hellenic Research Foundation	EKT/NHRF	http://www.ekt.gr/
	Open Access Publishing in European Networks	OAPEN	https://www.oapen.org/home
FR	CNRS (OpenEdition and Huma-Num)	CNRS	http://www.cnrs.fr/en
PL	The Institute of Literary Research of the Polish Academy of Sciences	IBL PAN	https://ibl.waw.pl/
PT	University of Coimbra	UC	https://digitalis.uc.pt/en
IT	University of Turin	UniTo	https://en.unito.it/
HR	University of Zadar	UniZd	https://www.unizd.hr/cimmar/involved-departments/department-of-information-sciences
CA	Public Knowledge Program	PKP	https://pkp.sfu.ca/
BR	Scientific Electronic Library Online	SciELO	https://scielo.org/en/
EU	Association of European University Presses (AEUP)	AEUP	https://www.aeup.eu/
IT	Associazione Italiana per la promozione della scienza aperta (AISA)	AISA	https://aisa.sp.unipi.it/about-aisa/
IT	Conferenza dei Rettori delle Universita Italiane/The Conference of Italian Univesity Rectors	CRUI	https://www.crui.it/crui-english.html

Data Harvesting Best Practices Document for Data Providers

PT	Faculty of Social Sciences and Humanities/Faculdade de Ciências Sociais e Humanas	FCSH	https://www.fcsh.unl.pt/
DE	Georg-August-University Göttingen		http://www.uni-goettingen.de/
FR	Huma-Num	HN	https://www.huma-num.fr/
US	Hypothesis		https://web.hypothes.is/
IT	Italian National Research Council	CNR	https://www.cnr.it/en
DE	Knowledge Unlatched	KU	https://www.knowledgeunlatched.org/
UK	KU Research (UK)		https://www.knowledgeunlatched.org/
IT	Lexis	Lexis	http://www.lexis.srl/
NL	Linguistics in Open Access	LingOA	https://www.lingoa.eu/
IT	Napoli University Federico II	UNINA	http://www.unina.it/en_GB/home
UK	Open Books Publishers	OBP	https://www.openbookpublishers.com/
UK	Open Library of Humanities	OLH	https://www.openlibhums.org/
FR	OpenEdition	OE	https://www.openedition.org
NL	Quality Open Access Market	QOAM	https://www.qoam.eu/
IT	Roma Tre University	ROMATRE	http://romatrepress.uniroma3.it/
SE	Stockholm University Library	Stockholm University Library	https://www.stockholmuniversitypress.se/

LX	The Luxembourg Centre for Contemporary and Digital History (C²DH)	C²DH	https://www.c2dh.uni.lu/
SI	The Research Center of the Slovenian Academy of Sciences and Arts / Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti	ZRC SAZU	https://www.zrc-sazu.si/en
UK	Ubiquity Press	Ubiquity Press	https://www.ubiquitypress.com/
UCL	UCL Press	UCL Press	https://www.uclpress.co.uk/
NO	UiT The Arctic University of Norway	UiT	https://en.uit.no/startsida
IT	University Ca'Foscari Venice	Ca' Foscari University of Venice	https://www.unive.it/pag/13526/
PT	University Institute of Lisbon/Instituto Universitário de Lisboa	ISCTE-IUL	https://www.iscte-iul.pt
BE	University of Liège	ULiège	https://www.uliege.be/cms/c_8699436/en/uliege
SI	University of Ljubljana, Faculty of Arts (SI)	University of Ljubljana, Faculty of Arts	http://www.ff.uni-lj.si/an/
IT	University of Milan	UNIMI	https://www.unimi.it/en
HR	University of Zagreb, University Computing Centre	SRCE	https://www.srce.unizg.hr/en/

B.   Annex II: TRIPLE general architecture

Data Harvesting Best Practices Document for Data Providers