GoTriple Content Providers Handbook

V2.0, April 2024

Table of content

Purpose of this document

GoTriple is a discovery service for scientific outputs in the Social Sciences and Humanities area. It is one of the services of the European Research Infrastructure OPERAS.
The GoTriple platform allows to search through publications, datasets, researchers' profiles, and research projects. Publications and datasets metadata are collected from both content aggregators and content providers (see Glossary).

This handbook is addressed to the **GoTriple content providers**. It presents the policies, technical requirements, and supporting actions enabling the providers' content acquisition and processing by the GoTriple platform. It informs them as well about other aggregators' policies.

GoTriple Policies

Scope

Scientific fields: GoTriple collects metadata of contents in the **Social Sciences and Humanities (SSH)** field (see Content types section). The content providers can be domain specific or not. In case they are not dedicated to SSH, the content providers are responsible for selecting SSH contents in their collections.

Data definition: GoTriple only collects **metadata of the contents**. The platform does not collect or store the contents of the providers.

Perimeter: GoTriple operates in the **European area** and provides enrichments for a limited number of European languages. Therefore, the platform mainly collects data from European providers and in European languages. However, it can also collect data from other geographic and linguistic areas.

Openness: As a service of OPERAS, the Research Infrastructure dedicated to open scholarly communication in the SSH, GoTriple promotes open access to the contents. **Openness** applies to datasets (open data) and publications (open access). The platform however collects open and not open access contents. The metadata of the contents need to be freely accessible and reusable.

Providers type: GoTriple providers can be of any size and provide large or small amount of contents in one or many SSH fields. GoTriple works with major aggregators, but also facilitates data acquisition from **small repositories or publishers**, like for instance diamond journals.

GoTriple supported languages: GoTriple provides enrichments of the metadata in the following languages: Croatian (HR), English (EN), French (FR), German (DE), Greek (EL), Italian (IT), Polish (PL), Portuguese (PT), Slovenian (SL), Spanish (ES), Ukrainian (UK). The platform however accepts contents in any language.

Content types

This handbook only considers the data collected on scientific resources. Data on researchers' profiles and research projects are collected through distinct automated processes.
In GoTriple, the resources are publications and datasets, which are all named "documents".
A **document** is the information asset, i.e. a series of metadata, of a unique and deduplicated resource (see Glossary).

The list of content types available on GoTriple is based on a subset of the COAR list of types (see Annexe).

On GoTriple, **publications** are any type of text or material related to the SSH research environment, from articles or thesis, to reports or learning material. As for the **datasets**, they are a set of organized research data. In the context of GoTriple, these resources are only indexed at the level of the collection. Each single file of the datasets is therefore not indexed. The level of collection harvestable by GoTriple will be assessed for each data source.

GoTriple Requirements

In order to have their contents collected by GoTriple, the providers essentially have to respect two requirements: providing an *access to their metadata* and providing *metadata compliant* with the platform's data model.

The providers can have their contents indexed on GoTriple in three ways: - by following the GoTriple recommendations for the technical set-up and the data model, i.e. using the OAI-PMH protocol and the DC metadata standard. - by sending a metadata file to the GoTriple administrators using the OpenAIRE format (see Annexe). - by having their contents on one of the aggregators of GoTriple (currently: DOAB, DOAJ, Isidore and OpenAIRE): in this case, the content will be automatically collected by GoTriple.

The first option ensures that the metadata is easily processable and can be regularly updated in an automated way.

The second option allows to provide metadata compliant with the GoTriple data model without using the OAI-PMH protocol. This solution however hinders automated data acquisition and updates.

The third option does not require any action from the providers if their content is already indexed by the listed aggregators, and if not, requires them to follow the aggregators specific guidelines (see Aggregators section).

The following two sections specify the technical set-up enabling GoTriple to automatically access the metadata and the model that these metadata should follow.

Technical set-up

GoTriple collects data by using the Open Archive Initiative - Protocol for Metadata Harvesting (OAI-PMH). The OAI-PMH protocol was developed in 1999 as part of the Open

Archives Initiative. It allows content providers to expose their metadata on the Web in a structured format and to make it available for harvesters. The repository is set-up by the provider and contains sets of metadata.

The metadata has to be represented according to the simple DublinCore standard at least, which uses `dc` elements that are described in the Dublin Core Metadata Element Set. A richer version of DC also exists, which uses `dcterms` elements and is described in the DCMI Metadata Terms. This version contains both simple Dublin Core (DC) and qualified Dublin Core (QDC). The OAI-PMH allows harvesters like GoTriple to then collect the structured metadata. DC and QDC can be expressed in XML, HTML, XHTML or JSON files, although GoTriple only collects `dc` and `dcterms` elements in XML format.

GoTriple also supports the harvesting of OAI-PMH endpoints that expose metadata in the Europeana Data Model (EDM), in the XOAI format and in the extension of Dublin Core used by the BASE aggregator (base_dc).

The OAI-PMH repository can be set up by the content provider with little development investment. Some tools and services for data repositories or publishers can contain an OAI-PMH module, either as a built-in feature or a plugin (e.g. DSpace, Eprints, OJS).
A list of open source tools to set up an OAI-PMH repository is available on the website of the Open Archive Initiative.

In the case of metadata provided through dedicated files, the publisher must periodically provide to GoTriple administrators a file export of their publications' metadata in the OpenAIRE format (see Annexe). The import procedure is not therefore completely automated and needs a certain number of human interactions. It is requested to provide at least two full export files per year per publisher via some file transfer method (http, ftp, cloud sharing, etc).

GoTriple data model

In order to ensure high semantic expressivity and address flexibility needs, the TRIPLE data model is based on the schema.org ontology, which is maintained by a World Wide Web Consortium (W3C) community. The ontology allows to handle the metadata of documents, but also of profiles and projects.

When it is collected through OAI-PMH, the metadata of the documents need to be compliant with DublinCore, simple or qualified, ie using `dc` or `dcterms` elements. When metadata is collected through formatted files, it is possible to use other schemas, like the aforementioned OpenAIRE format. Examples of the file both in DC and in OpenAIRE format are reproduced in Annexe).

Below, we describe the current TRIPLE data model for documents, specifying the level of priority, the corresponding `dc` and `dcterms` elements, and their expression in the TRIPLE data model.

Priority	Description	DublinCore	Triple data model
Mandatory	Creator of the resource	dcterms:creator, dc:creator	schema:author
Mandatory	Identifier of the resource	dcterms:identifier, dc:identifier	schema:identifier
Mandatory	Title of the resource	dcterms:title, dc:title	schema:headline
Recommended	Abstract	dcterms:description, dc:description, dcterms:abstract	schema:abstract
Recommended	Access rights to the resource	dcterms:rights, dc:rights	schema:conditionsOfAccess
Recommended	Date of publication or creation	dcterms:date, dc:date, dcterms:issued, dcterms:created, dcterms:available	schema:datePublished
Recommended	Keywords	dcterms:subject, dc:subject	schema:keywords
Recommended	Language of the resource	dcterms:language, dc:language	schema:inLanguage
Recommended	License	dcterms:license OR dcterms:rights, dc:rights	schema:license
Recommended	Publisher of the resource	dcterms:publisher, dc:publisher	schema:publisher
Recommended	Type of the resource	dcterms:type, dc:type	schema:additionalType
Recommended	URL of the landing page	dcterms:identifier dc:identifier	schema:mainEntityOfPage
Recommended	URL of the resource	dcterms:identifier dc:identifier	schema:url
Recommended	URL of the source (e.g. URL of a publishing platform)	dcterms:source, dc:source	schema:isBasedOnURL
Optional	Contributor to the resource's creation	dcterms:contributor, dc:contributor	schema:contributor
Optional	Format of the resource	dcterms:format, dc:format	schema:encodingFormat

Priority	Description	DublinCore	Triple data model
Optional	Information on the source (e.g. journal issue)	dcterms:source, dc:source	schema:mentions
Optional	Temporal coverage of the resource	dc:coverage, dcterms:coverage	schema:temporalCoverage
Optional	Spatial coverage of the resource	dcterms:spatial	schema:spatialCoverage

Priority.

The *mandatory* elements are necessary for the platform to process the metadata. The *recommended* elements increase both the findability of the contents and the quality of the automated processes run by the platform. The *optional* elements can provide additional information useful for the users.

DublinCore elements.

The simple DC elements are introduced by the `dc` namespace, the qualified DC elements are introduced by the `dcterms` namespace. While it is still technically possible to use `dc` elements, it is preferable to use `dcterms` elements, which allow for a more detailed description of the resource.
In some cases, it is possible to use one or many DC elements to describe an aspect of the resource: the date of the resource can be described through `dcterms:date` or through the more accurate `dcterms:created` and `dcterms:available`.

TRIPLE data model.

The elements of the TRIPLE data model in the schema.org ontology are reported only for information: these are handled only by the GoTriple platform, not by the content providers.
Some of the DC and QDC elements are processed by the platform. This is especially the case for "URL of the landing page" and "URL of the resource", which are automatically determined from the content of `dcterms:identifier` or `dc:identifier`. The TRIPLE data model also contains a few other elements for documents that are created through the analysis of the metadata files.

Best practices for contents' visibility

Metadata quality

While only three metadata elements are technically mandatory on GoTriple, richer metadata improves the processing by the information systems and therefore increases the visibility of the contents. However, some of the metadata elements require more accurate management in order to fully exploit the potentialities of the DublinCore standard. We list

below a few hints able to improve the metadata quality in the context of GoTriple, but also in the context of other aggregators, like OpenAIRE, DOAJ, DOAB or BASE.

Priority	Description	Hints and comments
Mandatory	Creator of the resource	Can contain one or many creators of the resource and can be individuals or organizations. On GoTriple, person names undergo a normalization process able to improve the filtering.
Mandatory	Identifier of the resource	Can contain one or many identifiers of different types. Identifiers are non semantic strings of characters uniquely identifying a resource. They should belong to a well-known identification system (e.g. ISBN, DOI, handle.net, etc.). In the digital context, the more important identifier is the Persistent Identifier (PID), which ensures the identification of the resource throughout the various digital locations. Persistent identifiers include among others: DOI from Datacite or Crossref, handles from handle.net.Identifiers should be provided as HTTP links and can be specified through dedicated encoding schemes accepted by the DC standard (e.g. URI).
Mandatory	Title of the resource	Titles are used for automated enrichments on GoTriple. They shouldn't be or contain a file name.
Recommended	Abstract	The `dcterms:description` can be more extended than the `dcterms:abstract`, or contain an abstract. On GoTriple, abstracts are used for the automated semantic classification and annotation.
Recommended	Access rights to the resource	On GoTriple, this information is retrieved exclusively from `dc:rights` or `dcterms:rights` elements. Note that a specific `dcterms:accessRights` also exists. Can contain free text information specifically about the access to the resource. As recommended also by OpenAIRE, it is possible to specify the access type in a normalized way through the COAR access rights types: embargoed access; metadata only access; open access; restricted access. Access information can be complemented with licensing information.
Recommended	Date of publication or creation	Without more precise information, the `dcterms:date` or `dc:date` element will be interpreted on GoTriple as "resource's first release date". Although the `date` element is normalized on GoTriple, it is preferable to use standardized date formats, like for instance ISO 8601. Date or period related not to the resource, but to its content, should be indicated in `dcterms:coverage`.

Priority	Description	Hints and comments
Recommended	Keywords	Can contain one or many keywords describing the content of the resource. In DC, the keywords language can be specified using an xml:lang attribute. The XML specification prescribes the usage of language identifiers as defined by IETF BCP 47 for values of this attribute. In the context of GoTriple it is mandatory to usegeneral, using the ISO 639-1 two-letter code.
Recommended	Language of the resource	Describes the language in which the resource is expressed. Like for keywords, for GoTriple, it is mandatory to use the ISO 639-1 two-letters code.
Recommended	License	A legal document indicating how the resource can be accessed and used. In QDC, there is a specific element for the licensing information: `dcterms:license`, but it is also possible to use `dcterms:rights`, `dc:rights` as an alternative. This element can also contain information about copyright and intellectual property rights. While a license can be a free text, it is preferable to use standardized licenses: they are easier to understand for humans and can facilitate machine-readability. In the context of open science, especially, it is recommended to use well-spread open licenses, for example Creative Commons (CC) licenses. The CC licenses allow to indicate an URL and to indicate the license type in a simple way. See the list of licenses supported by GoTriple in Annexe.
Recommended	Publisher of the resource	An entity responsible for making the resource available. Can be a person, an organization, like a publishing company or a service, like a data archive. The `publisher` element describes the resource and its production, not the creator and its affiliations.
Recommended	Type of the resource	The type of the resources should refer to a well-spread taxonomy, like the aforementioned COAR list of types, or the subset of COAR types used by GoTriple and listed in Annexe.
Recommended	URL of the landing page	The URL of the landing page can be indicated as a specific `dcterms:identifier` or `dc:identifier` element using the URI scheme. Should contain an HTTP link without a .pdf extension
Recommended	URL of the resource	Like the URL of the landing page, the URL of the resource itself can be indicated as a specific `dcterms:identifier` or `dc:identifier` element using the URI encoding scheme. In GoTriple, the URLs listed

Priority	Description	Hints and comments
		in the `identifier` elements containing a .pdf extension are used to create the direct link to the full text.
Recommended	URL of the source	A related resource from which the described resource is derived. In GoTriple, `dcterms:source` and `dc:source` elements are used to refer to the publishing platform or data repository if it contains an HTTP link without a .pdf extension.
Optional	Contributor to the resource's creation	An entity responsible for making contributions to the resource. Other than the entities that have contributed to the creation of the resource (e.g. a data scientist for a dataset, an editor for a publication), the `contributor` element can be used to list the organizations that have made the creation possible.
Optional	Format of the resource	The file format, physical medium, or dimensions of the resource. Recommended practice is to use a controlled vocabulary where available. For example, for file formats one could use the list of MIME Internet media types
Optional	Information on the source	This element is used for instance to indicate the relation of an article with a specific journal issue.
Optional	Temporal coverage of the resource	In the DublinCore standard, `dc:coverage` and `dcterms:coverage` contain information about temporal and spatial coverage. In GoTriple, these elements are used only for the temporal coverage. Spatial coverage should be indicated in `dcterms:spatial` .
Optional	Spatial coverage of the resource	Contains free or standardized text about the spatial area considered by the resource. The element is only available in QDC: `dcterms:spatial` .

FAIR principles

The FAIR principles emerged in 2016 from an interdisciplinary group of research data experts. The acronym FAIR refers to four guiding principles for digital data management: making the data Findable, Accessible, Interoperable, and Reusable. The FAIR principles address the need for a common understanding of data management good practices able to facilitate data sharing and reuse. Although the principles mainly consider technical aspects, they allow, as principles, to adapt the concrete implementations to specific contexts. In particular, they can apply to any research digital object: datasets, publications, software, etc.

The four ground principles are further described in a set of fifteen principles. GOFAIR, the organization supporting the FAIR principles adoption, gives detailed information about the fifteen FAIR principles. The OPERAS Special Interest Group on "Common standards and FAIR principles" also provided an overview in its 2021 White paper.

The FAIR principles are a useful tool to manage digital data in a way that facilitates both human and machine operations. They have been used to build the TRIPLE data model and are now at the core of all major aggregators practices.

The **Findability** principle relies mainly on the use of persistent identifiers and rich descriptive metadata. The metadata should give information about the resource, like: creator, title, persistent identifier, publisher, publication date, abstract and keywords. A counterexample is a corpus stored on a USB device with descriptive information only in the file name.

The **Accessibility** principle relies on the use of open, free and documented protocols, such as HTTP, OAI-PMH, FTP, even if they are combined with authentication processes. Accessibility is further improved if metadata gives information about the conditions of access. A counterexample is the data exchanged through emails on individual request.

The **Interoperability** principle relies on the use of standard representation of the data, like the DublinCore schema aforementioned. Interoperability can also be reached thanks to documented controlled vocabularies shared within a broad community. A counterexample is a dataset described according to an individual and not documented vocabulary.

The **Reusability** principle relies on clear licensing information, as liberal as possible, preferably standard and recorded in the metadata. It should be completed with clear provenance information, which allows to better assess the reusability possibilities. A counterexample, without mentioning the lack of any license, is a license containing unclear conditions of use in a separate PDF file.

The TRIPLE data model follows the main aspects of the FAIR principles and this handbook will help you to ensure that your content is Findable, Accessible, Interoperable, and Reusable.

Aggregators

GoTriple harvests metadata from major European aggregators and the TRIPLE data model is globally compliant with their own data models. Your content can therefore appear on GoTriple if it is already indexed by our current partner aggregators: DOAB, DOAJ, Isidore and OpenAIRE. Furthermore, if the content provider doesn't have an OAI-PMH repository, GoTriple can ingest metadata files formatted according to the OpenAIRE format (see Annexe).
However, with respect to the TRIPLE data model, each aggregator may have additional requirements. We list below some information about these requirements for the main aggregators that are useful for the SSH research community.

Here is the list of aggregators currently harvested by GoTriple:

- DOAB (*harvested by GoTriple*).

DOAB is a community-driven discovery service that indexes and provides access to scholarly, peer-reviewed open access books.

The current main requirements to provide content in DOAB are twofold:

1/ Academic books in DOAB shall be available under an open access licence (such as a Creative Commons licence).

2/ Academic books in DOAB shall be subjected to independent and external peer review prior to public.

From a technical point of view, the requirements are specified in a separate document. Although the TRIPLE data model is globally compliant with the DOAB data model, the latter contains information specific to monographs. The DOAB also handles an OAI-PMH repository, but the content provider can send the metadata through a form or a file, without using DublinCore or an OAI-PMH repository.

- DOAJ (*harvested by GoTriple*).

As their website reads, DOAJ is an independent index containing almost 17 500 peer-reviewed, open access journals. The DOAJ covers all areas of science, technology, medicine, social sciences, arts and humanities, with open access journals from all countries and in all languages.

The criteria for the journals are the following:

1/ The journal must be actively publishing scholarly research.

2/ Publish in any research subject area.

3/ Should publish at least 5 research articles per year.

4/ Its primary target audience should be researchers or practitioners.

For newly launched journals, an additional requirement applies: it must demonstrate a publishing history of more than one year, or have published at least 10 articles.
The DOAJ doesn't require to have an OAI-PMH repository. There are various ways to upload metadata about articles to DOAJ. It is possible to send JSON files via their API (see documentation). It is also possible, once the journal has been accepted to use a dedicated space to send metadata in an XML file or to enter it manually. The DOAJ provides an application guide for the journals publishers. More details about the DOAJ's data model for journals and articles can be found on this page. DOAJ is harvested by the main aggregators, including GoTriple. As the DOAJ does not require to set up an OAI repository, it offers therefore a good solution for content providers who do not have an OAI repository available.

- Isidore (*harvested by GoTriple*).

Isidore is a French platform dedicated to social sciences and humanities scientific outputs. It contains over than 10 millions documents, ranging from publications to data (full list of types). Isidore has the specificity to offer high quality automated enrichments in three languages (ENG, FRA, SPA) aligned with major authority databases (e.g. RAMEAU, LCSH, BNE) and exposed in a structured way compliant with semantic web requirements. Isidore served as a basis for the building of GoTriple.

Isidore relies on the OAI-PMH protocol to retrieve data. Although enrichments are proposed only in three languages, the platform can harvest contents in any language. At present, it collects metadata from a variety of publishing platforms, archives, and repositories in Europe and beyond. The data sources are listed in the documentation (French only).
In order to have contents harvested by Isidore, it is required to have a valid OAI-PMH endpoint and provide simple or qualified DublinCore metadata. The guidelines for metadata management provided by the DublinCore Metadata Innovation group are applicable for the harvesting by Isidore.

Any repository or publisher interested in being harvested by Isidore should contact directly the administrators of the platform using this email address: isidore-sources(at)huma-num.fr.

- BASE (*harvested by GoTriple*).

BASE (Bielefeld Academic Search Engine) is one of the world's largest search engines for academic web resources. Operated by Bielefeld University Library, BASE indexes metadata from a wide range of scholarly sources, including institutional repositories, digital collections, and open access journals. With over 300 million documents from more than 10,000 content providers, BASE offers a powerful and user-friendly interface for researchers to discover academic articles, theses, books, and other scholarly materials. The service prioritizes open access content and is freely accessible to the public at https://base-search.net/.

- OpenAIRE (*harvested by GoTriple*).

Among other services, OpenAIRE offers a discovery service for millions of research outputs: publications, datasets, software and other research products. Although mostly referencing contents produced in the European area, OpenAIRE also accepts contents from outside Europe.
GoTriple harvests from OpenAIRE a subset of publications and datasets in the SSH area.

OpenAIRE relies on the OAI-PMH protocol to retrieve data. The protocol allows OpenAIRE to harvest data from the zenodo repository and from the research data repositories listed by re3data. It is therefore possible to be harvested by OpenAIRE and indexed on their platform Explore, by having the content stored in one of these repositories.
It is also possible to become a content provider for OpenAIRE. Besides the requirement to have an operational OAI-PMH endpoint, becoming a content provider requires to respect these two main requirements:

1/ Having your repository registered on OpenDOAR.

2/ Following the OpenAIRE guidelines to establish your metadata.

The OpenAIRE metadata guidelines depends on the type of digital content:
- Guidelines for Literature, institutional, and thematic Repositories. - Guidelines for Data Archives. - Guidelines for Software Repository Managers. - Guidelines for Other Research Products.

More information about OpenAIRE's content policy can be found in the dedicated section of their website, especially the Terms of Use document which has to be reciprocally accepted by OpenAIRE and the content provider.

Process and support

Harvesting management system

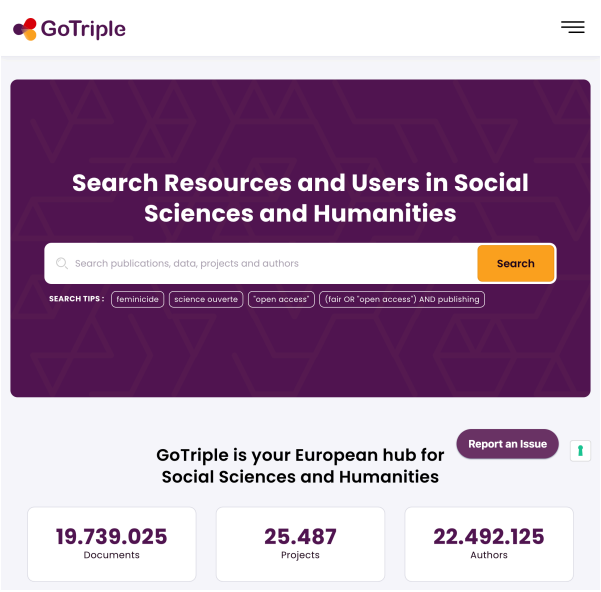
Content providers interested in being harvested by GoTriple should first check the GoTriple policies and requirements described in this document (see GoTriple policies and GoTriple requirements).

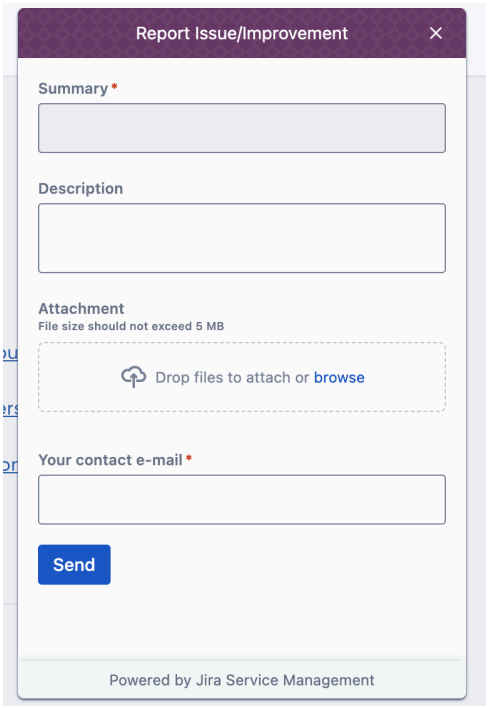
Providers fulfilling the GoTriple prerequisites, i.e. having installed an OAI-PMH repository with standard Dublin Core metadata available, or who would like to send database dumps structured according to the OpenAIRE or MARC21-XML formats should contact the GoTriple admin team (see section "How to get in touch" below).

How to get in touch

To propose your content in GoTriple you can get in touch with our team with the following ways:

- by sending a request via the "Report an issue" form on the GoTriple website.
- by email: write to gotriple@operas-eu.org
- by participating to the GoTriple community by registering to the Mattermost channel GoTriple/Community Discussion





Credits

The GoTriple platform and this handbook have been created through the project TRIPLE (Transforming Research Through Innovative Practices for Linked Interdisciplinary Exploration).
TRIPLE has received funding from the European Union's Horizon 2020 Research and Innovation action funding scheme INFRAEOSC-02-2019 "Prototyping new innovative services" (grant agreement #863420).

Annexe

Glossary

Aggregator
An organization that collects, manages, and disseminates the metadata of the scholarly resources' made available by various providers. The aggregator operates as a standardisation body of heterogeneous metadata, either by defining its own requirements, or by relying on existing standards for harvesting and dissemination.

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

Dataset
On GoTriple, corresponds to a collection of files produced within a research project for

analysis and processing. A dataset can contain files of various types (surveys, recordings, images, measures, etc.) and formats (video, audio, spreadsheets, etc.).

Document
Refers to the information asset related to a specific digital object; it is used to identify single scholarly resources such as publications and datasets. On GoTriple, a document corresponds to the set of metadata, collected or generated, describing a scholarly resource.

Harvester In the context of OAI-PMH, the Open Archive Initiative - Protocol for Metadata Harvesting, a harvester is the entity collecting automatically metadata exposed in an OAI-PMH repository. Harvesters can collect metadata from many providers, in which case the harvesters can be defined as aggregators.

Provider
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 metadata facilitating its dissemination, and acts as the primary dissemination body of the data and its metadata.

Publication On GoTriple, a publication is a textual object formatted for dissemination and produced in the context of research. A publication can be of various types (article, book, report, training, etc.) and formats (pdf, epub, presentation, etc.).

GoTriple list of content types

- article (COAR type)
- bibliography (COAR type)
- blog-post (COAR type)
- book (COAR type)
- book part (COAR type)
- conference (COAR type)
- dataset (COAR type)
- image (COAR type)
- learning-object (COAR type)
- manuscript (COAR type)
- report (COAR type)
- periodical (COAR type)
- preprint (COAR type)
- review (COAR type)
- software (COAR type)
- text (COAR type)
- thesis (COAR type)
- map (COAR type)
- other (COAR type)

List of licenses supported on GoTriple

GoTriple will provide normalization for a a limited set of widely used set of licenses. Here is the list of the supported licenses: - Cairn - Creative Commons - OpenSource licenses (including apache, gpl, bsd, mit licence) - CLARIN PUB, CLARIN ACA, CLARIN ACA-NC, CLARIN-RES, CLARIN RES-NC - Microsoft Public Licence - Microsoft Reciprocal Licence - Open Data (ODbL, Open Data Commons Open Database Licence) - META-SHARE No Redistribution, META-SHARE NonCommercial NoRedistribution, META-SHARE Commercial No Redistribution For a Fee, META-SHARE Noncommercial No Redistribution For a Fee - ELRA licenses - other licenses will be marked as “other” if not recognized, or “undefined” if the license is not specified

Sample metadata files

Below are displayed two examples of the files used on the GoTriple platform: one in the XML Dublin Core format, one in the JSON OpenAIRE format.

XML DC example:

```
<record>
    <header
        xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
        <identifier>oai:doaj.org/article:eebc19b7f56c4c439b316061bffd423d</identifier>
        <datestamp>2022-12-21T23:20:41Z</datestamp>
        <setSpec>TENDOkhpc3Rvcnkgb2Ygc2Nob2xhcnNoaXAgYW5kIGxlYXJuaW5nLiBUaGUgaHVtYW5pdGllcw~~</setSpec>
        <setSpec>TENDOkVsZWN0cm9uaWMgY29tcHV0ZXJzLiBDb21wdXRlciBzY2llbmNl</setSpec>
    </header>
    <metadata
        xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
        <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
            <dc:title>Asymmetric Digital Collaboration and Collective Authorship: On Digital Genres and Writing
            Processes for 'CanLit Guides'</dc:title>
            <dc:identifier>1918-3666</dc:identifier>
            <dc:identifier>10.16995/dscn.28</dc:identifier>
            <dc:identifier>https://doaj.org/article/eebc19b7f56c4c439b316061bffd423d</dc:identifier>
            <dc:date>2016-03-01T00:00:00Z</dc:date>
            <dc:relation>https://www.digitalstudies.org//articles/28</dc:relation>
            <dc:relation>https://doaj.org/toc/1918-3666</dc:relation>
            <dc:description>This paper discusses the unique asymmetric collaboration process used at CanLit
            Guides in the first phase of its development. CanLit Guides began as a project to mobilize the
            massive digital archive (1959-2008) of the scholarly journal Canadian Literature. The Guides
            introduce undergraduate students to areas of scholarly and critical concern in the larger field of
            Canadian Literature and culture. The editors of Canadian Literature enabled graduate students to
            develop teamwork, research, teaching, and digital writing skills by employing them as developers,
            researchers, and writers. /.../ </dc:description>
            <dc:creator>Mike Borkent</dc:creator>
            <dc:creator>Jamie Paris</dc:creator>
            <dc:publisher>Open Library of Humanities</dc:publisher>
            <dc:type>article</dc:type>
            <dc:subject>CanLit Guides, workflow, digital genres, digital pedagogical project, collaboration,
            authorship, literature and culture</dc:subject>
            <dc:subject xsi:type="dcterms:LCC">History of scholarship and learning. The humanities</dc:subject>
            <dc:subject xsi:type="dcterms:LCC">AZ20-999</dc:subject>
            <dc:subject xsi:type="dcterms:LCC">Electronic computers. Computer science</dc:subject>
            <dc:subject xsi:type="dcterms:LCC">QA75.5-76.95</dc:subject>
            <dc:language>EN</dc:language>
            <dc:rights>OPEN</dc:rights>
            <dcterms:license>cc-by</dcterms:license>
            <dcterms:temporal>1999-01-01</dcterms:temporal>
            <dcterms:spatial>France</dcterms:spatial>
            <dc:source>Digital Studies (2016)</dc:source>
            <dcterms:contributor>Bureau Interdisciplinaire Landschapsanalyse</dcterms:contributor>
            <dcterms:format>application/pdf</dcterms:format>
        </oai_dc:dc>
    </metadata>
</record>
```

JSON OpenAIRE example:


```
{
  "author": [
    {
      "fullname": "Mike Borkent"
    },
    {
      "fullname": "Jamie Paris"
    }
  ],
  "bestaccessright": {
    "label": "OPEN"
  },
  "collectedfrom": [
    {
      "key": "10|driver______::bee53aa31dc2cbb538c10c2b65fa5824"
    },
    {
      "key": "10|openaire____::081b82f96300b6a6e3d282bad31cb6e2"
    },
    {
      "key": "10|openaire____::8ac8380272269217cb09a928c8caa993"
    },
    {
      "key": "10|openaire____::5f532a3fc4f1ea403f37070f59a7a53a"
    }
  ],
  "contributor": [
    "Bureau Interdisciplinaire Landschapsanalyse"
  ],
  "coverage": [
    "1999-01-01"
  ],
  "dateofcollection": "2022-12-21T23:20:41Z",
  "description": [
    "This paper discusses the unique asymmetric collaboration process used at CanLit Guides in the first phase
    of its development. CanLit Guides began as a project to mobilize the massive digital archive (1959-2008) of
    the scholarly journal Canadian Literature. The Guides introduce undergraduate students to areas of
    scholarly and critical concern in the larger field of Canadian Literature and culture. The editors of
    Canadian Literature enabled graduate students to develop teamwork, research, teaching, and digital writing
    skills by employing them as developers, researchers, and writers. The project supports open access,
    scholarly collaboration, and the creation of new digital genres. As the project evolved, however, it became
    clear that getting a team of scholars to work on a hierarchized, or what we call \"asymmetric,\"
    collaboration between the editors and the graduate students, is particularly difficult, and can lead to
    issues of doneness and sprawl. Producing a collaborative and democratic workflow process enabled us to
    write a robust collection of guides in innovative digital genres. This paper pays particular attention to
    issues of authorship that come up with any collaborative digital writing project, and it discusses the
    complexities of the graduate student experience of working on a digital pedagogical development team.
    Cet article discute du processus unique de collaboration asymetrique qui a ete utilise dans les guides sur
```

```
"format": [
  "application/pdf"
],
"id": "50|dedup_wf_001::fb93f67c7220dc13b3e4dc7cb39aefab",
"instance": [
  {
    "collectedfrom": {
      "key": "10|driver______::bee53aa31dc2cbb538c10c2b65fa5824"
    },
    "hostedby": {
      "key": "10|doajarticles::d1c58936cca4fc19deb1329841774135"
    },
    "type": "Article",
    "url": [
      "https://www.digitalstudies.org//articles/28"
    ]
  },
  {
    "collectedfrom": {
      "key": "10|openaire____::081b82f96300b6a6e3d282bad31cb6e2"
    },
    "hostedby": {
      "key": "10|doajarticles::d1c58936cca4fc19deb1329841774135"
    },
    "license": "http://creativecommons.org/licenses/by/4.0",
    "type": "Article",
    "url": [
      "http://dx.doi.org/10.16995/dscn.28"
    ]
  },
  {
    "collectedfrom": {
      "key": "10|openaire____::8ac8380272269217cb09a928c8caa993"
    },
    "hostedby": {
      "key": "10|doajarticles::d1c58936cca4fc19deb1329841774135"
    },
    "license": "cc-by",
    "type": "Article",
    "url": [
      "https://doi.org/10.16995/dscn.28"
    ]
  },
  {
    "collectedfrom": {
      "key": "10|openaire____::5f532a3fc4f1ea403f37070f59a7a53a",
      "value": "Microsoft Academic Graph"
    },
```

```
"language": {
  "code": "eng"
},
"maintitle": "Asymmetric Digital Collaboration and Collective Authorship: On Digital Genres and Writing
Processes for 'CanLit Guides'",
"originalId": [
  "oai:doaj.org/article:eebc19b7f56c4c439b316061bffd423d",
  "10.16995/dscn.28",
  "2328671840"
]
"publicationdate": "2016-03-30",
"publisher": "Open Library of Humanities",
"subjects": [
  {
    "subject": {
      "scheme": "keyword",
      "value": "CanLit Guides, workflow, digital genres, digital pedagogical project, collaboration,
      authorship, literature and culture"
    }
  },
  {
    "subject": {
      "scheme": "lcsh",
      "value": "lcsh:History of scholarship and learning. The humanities"
    }
  },
  {
    "subject": {
      "scheme": "lcsh",
      "value": "lcsh:AZ20-999"
    }
  },
  {
    "subject": {
      "scheme": "lcsh",
      "value": "lcsh:Electronic computers. Computer science"
    }
  },
  {
    "subject": {
      "scheme": "lcsh",
      "value": "lcsh:QA75.5-76.95"
    }
  }
]
```


OpenAIRE JSON file format

The OpenAIRE JSON schemas for publications are available here https://zenodo.org/record/5799514#.YtAAYOxBw40. The specific schema used by GoTriple is: https://zenodo.org/record/5799514/files/result_schema.json?download=1.

In GoTriple we only consider the following subset of elements, which must be included in the dump produced for the import in our platform:

Description	JSON schema	Triple data model
Creator of the resource	author/fullname	schema:author
Identifier of the resource	id	schema:identifier
Title of the resource	maintitle	schema:headline
Abstract	description	schema:abstract
Access rights to the resource	bestaccessright/label	schema:conditionsOfAccess
Date of publication or creation	publicationdate	schema:datePublished
Keywords	subjects/subject/value (Only if subjects/subject/scheme=keyword)	schema:keywords
Language of the resource	language/code	schema:inLanguage
License	instance/license	schema:license
Publisher of the resource	publisher	schema:publisher
Type of the resource	instance/type	schema:additionalType
URL of the landing page	instance/url (it must start with http and must not end with .pdf)	schema:mainEntityOfPage
URL of the resource	instance/url (it must start with http and end with .pdf)	schema:url
URL of the source (e.g. URL of a publishing platform)	we consider the element: originalId (it must start with http)	schema:isBasedOnURL
Contributor to the resource's creation	contributor	schema:contributor
Format of the resource	format	schema:encodingFormat
Information on the source (e.g. journal issue)	we consider the element: collectedfrom/key (it must not start with http)	schema:mentions

Description	JSON schema	Triple data model
Temporal coverage of the resource	null	schema:temporalCoverage
Spatial coverage of the resource	coverage	schema:spatialCoverage