

	 <p>Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration</p>
[DECEMBER 2022]	Advancing Open Scholarship
	<b>D2.6 – REPORT ON GLOBAL DATA RETRIEVAL</b> Version 1.0 – Final/PUBLIC
	H2020-INFRAEOSC-2019 Grant Agreement 863420

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

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

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



# REPORT ON GLOBAL DATA RETRIEVAL

---

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	D2.6
Contractual Delivery Date: 31/12/2022	Actual Delivery Date: 31/12/2022
Nature: Report	Version: 1.0 Final
Dissemination level	PU

## Revision History

Version	Created/Modifier	Comments
0.0	Veronika Fedotova (OPERAS)	Notes for the deliverable
0.1	Luca De Santis (NET7), Arnaud Gingold (OPERAS-AMU), Marta Błaszczńska, Maciej Maryl (IBL PAN)	First full version
0.2	Francesca Di Donato (CNR), Haris Georgiadis (EKT), Tomasz Umerle, Cezary Rosiński, Nikodem Wołczuk, Mateusz Franczak (IBL PAN), Taina Jääskeläinen (CESSDA), Emilie Blotière (OPERAS)	Revisions and comments
1	Luca De Santis, Arnaud Gingold, Marta Błaszczńska, Maciej Maryl	First released version

## Table of Contents

Summary	4
1. Introduction	5
2. Source selection and description	5
2.1 Policies	5
2.2 Selection process	6
2.3 Data acquisition	8
2.4 GoTriple content providers handbook	11
Overview:	11
3. Harvesting process	11
3.1 Harvesting process in the context of the project	11
3.2 Harvesting management tool	16
Conclusions	28
Lessons learned	28
Next steps	29

## List of Figures

- Figure 1 - SCRE software architecture (p. 12)
- Figure 2 - Publications data flow (p. 13)
- Figure 3 - Projects data flow (p. 13)
- Figure 4 -The original UI of the SCRE web application (p. 17)
- Figure 5 - Registration form for Data Providers in the HMS (p. 19)
- Figure 6 - Verification of a Data Provider request for registration (p. 20)
- Figure 7 - Data Providers home page in the HMS (p. 21)
- Figure 8 - Proposing a source by a Data Provider: specifying the OAI-PMH endpoint (p. 22)
- Figure 9 - Proposing a source by a Data Provider: verifying a preview of the fetched results (p. 23)
- Figure 10 - Proposing a source by a Data Provider: detailed view of a fetched result (p. 24)
- Figure 11 - Proposing a source by a Data Provider: final confirmation step (p. 24)
- Figure 12 - Content Editors view to accept or discard a proposed source (p. 25)
- Figure 13 - Preview of a source (p. 26)
- Figure 14 - Statistics of a source and of a flow (p. 27)

## List of Tables

- Table 1. Acquisition completed (pp. 9-10)
- Table 2. Acquisition on-going (p. 10)

# Acronyms

---

COESO	Collaborative Engagement on Societal Issues
HMS	Harvesting Management System
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
SSH	Social Sciences and Humanities
TRIPLE	Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration

## Summary

---

This deliverable summarises the different stages and processes of data acquisition for GoTriple, a platform created in the context of the TRIPLE project. It also presents the reasons for adopting them and the challenges and opportunities identified by the team. When relevant, the report points to other TRIPLE deliverables covering different aspects of the data retrieval process and the rationale behind selected solutions.

Firstly, the processes related to source selection and descriptions are presented. GoTriple collects metadata of documents in the social sciences and humanities (SSH) field in any language, with a focus on open access content and the European Research Area (ERA). It collects data from both aggregators and providers. Completed and ongoing (status: December 2022) data acquisition processes are summarised, together with the types of processes selected for them. The report also explains the purpose of the GoTriple content providers handbook that is currently set up and will constitute the support service for GoTriple providers.

Secondly, the harvesting process, including the Harvesting Management Support tool, are described. SCRE (Semantic Content Retrieval Engine), a dedicated platform for data ingestion and curation, has been developed as part of the project to process data by using a pipeline approach. Concrete steps that the tool offers to the user are described and illustrated with screenshots.

Lastly, the deliverable provides reflections on the data retrieval activities and choices that have taken place during the project and directions in which next steps of GoTriple development should be taken.

# 1. Introduction

The [GoTriple platform](#) has been created as a result of the [TRIPLE](#) project (“Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration”), financed under the Horizon 2020 Framework Programme (2019-2023). The platform provides an access point to social sciences and humanities (SSH) content harvested from multicultural and multilingual sources, with 10 European languages supported.<sup>1</sup>

The purpose of the D2.6 Report on Global Data Retrieval deliverable is to provide an overview of the data acquisition process, summing up the work conducted in this area for GoTriple. First, it provides a description of sources and their selection. Second, it outlines and explains the harvesting process. It also presents the Harvesting Management Support tool.

Different aspects of the data retrieval process and the rationale behind selected solutions are covered in more detail in the following deliverables:

- D2.1 “[Data acquisition plan](#)”;
- D2.2 “[Data harvesting best practices for data providers](#)”;
- D2.3 “[Report on Machine Learning](#)”;
- D2.4 “Report on Identification and Creation of New Vocabularies”;
- D2.5 “[Report on data enrichment](#)”.

## 2. Source selection and description

### 2.1 Policies

GoTriple collects metadata of documents in the SSH field in any language, with a focus on open access content and the European Research Area (ERA). Within GoTriple, documents are publications or datasets; the platform search filters allow to increase the visibility of open access documents and of publications with full text available. Detailed policies for data sources’ selection have been established and integrated into the handbook for content providers (see section “Harvesting management and support” of the handbook, to be released in March 2023).

The handbook describes the GoTriple policies as follows:

- **Scientific fields:** GoTriple collects metadata of content in the Social Sciences and Humanities (SSH) field (see Content types section of the handbook). The content

---

<sup>1</sup> Croatian (HR), English (EN), French (FR), German (DE), Greek (EL), Italian (IT), Polish (PL), Portuguese (PT), Spanish (ES), Ukrainian (UK).

providers can be domain-specific or not. In case they are not dedicated to SSH, the content providers are responsible for selecting SSH content in their collections.

- **Data definition:** GoTriple collects only metadata. The platform does not collect nor store the actual documents.
- **Scope:** GoTriple operates in the European Research Area and provides enrichment for a limited number of European languages. Therefore, the platform mainly collects data from European providers and in European languages. However, it is capable of collecting data from other geographic and linguistic areas.
- **Openness:** As a service of OPERAS, the ESFRI-roadmap Research Infrastructure dedicated to open scholarly communication in the SSH, GoTriple promotes open access to the content. Openness applies to datasets (open data) and publications (open access). The platform however collects metadata of both open and restricted access content. The metadata of the content needs to be freely accessible and reusable.
- **Providers type:** GoTriple providers can be of any size and provide varying volumes of content in one or many SSH fields. GoTriple mainly works with 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), Spanish (ES), Ukrainian (UK). The platform however accepts content in any language.

Upon this basis, aggregators and providers have been identified and selected as potential sources for GoTriple.

## 2.2 Selection process

The selection process started with the stock-taking of potential sources for GoTriple, both within and outside the consortium. The consortium first identified a list of potentially compliant aggregators and providers in the European Research Area.

As the aggregators often are multidisciplinary and use a specific data model, the selection process requires more in-depth discussions. Therefore, in the first phase of the project

(during 2021 Q1-Q2) the project team conducted a series of meetings with such aggregators as DOAB<sup>2</sup>, DOAJ<sup>3</sup>, Europeana<sup>4</sup>, Istex<sup>5</sup>, NARCIS<sup>6</sup>, OpenAIRE<sup>7</sup>.

These meetings were focused on compliance with GoTriple policies and the technical implementation of the data acquisition. The discussion about compliance mainly considered the identification of SSH content in the aggregators' collections and the amount of openly accessible resources. The discussion about the implementation resulted in envisioning three key acquisition pathways: direct import through OAI-PMH (providers or aggregators), direct import through database dumps (providers or aggregators), indirect import of providers' data through aggregators (e.g. Narcis harvested through Openaire).

After this first round of meetings, it was possible to expand the search for data sources to smaller providers. A campaign launched within the consortium and its network (during 2021 Q3) led to the identification of the following new sources: Biblioteka Nauki<sup>8</sup> (PL), ZRC SAZU<sup>9</sup> (SL), Econstor<sup>10</sup> (DE). These were added to providers already represented in the consortium: EKT publishing, National Archive of PhD thesis, SearchCulture.gr collections (GR), Pombalina (University of Coimbra - PT).

While the main obstacles to data source validation were the selection of SSH content, compliance with international metadata standards, and level of openness, the integration of the validated sources allowed to refine and improve the data acquisition process.

The selection process led to the discarding of the following sources: Istex (FR), NARCIS (NL) and FBC (PL). Istex is a French service whose access is partially limited to researchers working in France, while its open-access content is already retrieved by GoTriple from other sources. NARCIS's content described with good quality metadata is harvested by OpenAIRE, therefore NARCIS's managers recommended collecting their metadata through this aggregator. It has also been envisioned to harvest metadata from FBC<sup>11</sup>, which is the Federation of Polish Digital Libraries. However, FBC does not support OAI-PMH for the extraction of its data, so automatic and repeatable extraction is not possible. Also, the filtering of SSH content from this robust resource is challenging.

---

<sup>2</sup> Directory of Open Access Books: <https://www.doabooks.org/>.

<sup>3</sup> Directory of Open Access Journals: <https://doaj.org/>.

<sup>4</sup> <https://www.europeana.eu/en>.

<sup>5</sup> Information Scientifique et Technique d'Excellence/Scientific and Technical Information of Excellence: <https://www.istex.fr/>.

<sup>6</sup> National Academic Research and Collaborations Information System (NL): <https://www.narcis.nl/>.

<sup>7</sup> <https://www.openaire.eu/>.

<sup>8</sup> Biblioteka Nauki/Library of Science: <https://bibliotekanauki.pl/>.

<sup>9</sup> Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti/Scientific Research Centre of the Slovenian Academy of Sciences and Arts: <https://www.zrc-sazu.si/en>.

<sup>10</sup> <https://www.econstor.eu/>.

<sup>11</sup> <https://fbc.pionier.net.pl/>.

All the other sources have been validated and their metadata acquisition was achieved either through the generic OAI-PMH solution adopted by GoTriple, or through more specific processes described further in this report.

## 2.3 Data acquisition

As described in the D2.1 “[Data acquisition plan](#)”<sup>12</sup> and the D2.2 “[Data harvesting best practices for data providers](#).”<sup>13</sup> GoTriple collects metadata from both aggregators and providers. The specific processes of data ingestion and metadata mappings are described in D2.5 “[Report on data enrichment](#)”. This section provides an overview of the process.

Aggregators are defined as secondary sources of data, while providers are primary sources of data. The volume and type of data provided by the aggregators often require a specific data acquisition process on GoTriple, like it is the case for OpenAIRE or Isidore, which necessitated specific connectors and mappings (see [D2.5](#) and further sections of this document).

The tables 1 and 2 summarize the acquisition status and content of the following sources:

- Acquisition completed:
  - DOAB: directory of open access books in all disciplines
  - DOAJ: directory of open access journals in all disciplines
  - Isidore: directory of scientific resources in the SSH
  - OpenAIRE: directory of scientific resources in all disciplines
  - Biblioteka Nauki: platform of open access publications in all disciplines
  - EKT publishing: platform of open access publications in all disciplines
  - EKT National Archive of PhD Theses
  - EKT Searchculture.gr
  - OAPEN: platform of open access publications in all disciplines
  - ZRC SAZU: platform of open access publications in all disciplines
  - CESSDA: platform of research data in the social sciences
- Acquisition on-going:
  - BASE: directory of scientific resources in all disciplines

---

<sup>12</sup> <https://doi.org/10.5281/zenodo.4311460>.

<sup>13</sup> <https://doi.org/10.5281/zenodo.4438650>.



- Europeana: platform of scientific resources about cultural heritage
- CLARIN centers: platforms of research data in linguistics
- Econstor: directory of open access publications in economics
- OpenEdition: platform of open access publications in the SSH
- Pombalina: platform of open access publications in the SSH

These acquisitions should be completed in 2023 Q1. In the case of BASE and Europeana, further assessments are required to establish a more precise timeline for ingestion.

**Table 1. Acquisition completed**

<b>Aggregators</b>					
<b>Name</b>	<b>Type</b>	<b>Documents</b>	<b>Schema</b>	<b>Process</b>	<b>Updates frequency</b>
DOAB	Publications	41 000	Dublin Core	OAI-PMH	Daily
DOAJ	Publications	1 M	Dublin Core	OAI-PMH	Daily
isidore	Mixed	3 M	schema.org	XML dump	6 months
OpenAIRE	Mixed	200 000	OpenAIRE	Json dump	6 months
EKT SearchCulture.gr	Mixed	96 206	EDM	OAI-PMH	Daily
<b>Providers</b>					
Biblioteka Nauki	Publications	250 000	Dublin Core	OAI-PMH	Daily
EKT publishing	Publications	25 000	Dublin Core	OAI-PMH	Daily
EKT National Archive of	Theses	11 756	Dublin Core	OAI-PMH	Daily

PhD Theses					
OAPEN	Publications	14 000	Dublin Core	OAI-PMH	Daily
ZRC-SAZU	Publications	16 400	Dublin Core	OAI-PMH	Daily

**Table 2. Acquisition on-going (status: December 2022)**

<b>Aggregators</b>					
<b>Name</b>	<b>Type</b>	<b>Documents</b>	<b>Schema</b>	<b>Process</b>	<b>Updates frequency</b>
BASE	Mixed	TBD	Dublin Core	search API / OAI-PMH	TBD
CESSDA	Data	19 000	Dublin Core	OAI-PMH	Daily
Europeana	Mixed	TBD	EDM	OAI-PMH	TBD
<b>Providers</b>					
CLARIN centers	Data	TBD	Dublin Core	OAI-PMH	Daily
Econstor	Publications	TBD	Dublin Core	OAI-PMH	Daily
OpenEdition	Publications	TBD	Dublin Core	OAI-PMH	Daily
Pombalina	Publications	3 000	Dublin Core	OAI-PMH	Daily

## 2.4 GoTriple content providers handbook

The handbook is a follow-up of D2.1 “[Data acquisition plan](#)” and the D2.2 “[Data harvesting best practices for data providers](#)” presented in a format more accessible for any data provider, even non expert in information technology. It summarizes the content of the technical deliverables, providing useful information and advice to become a GoTriple data

provider. The handbook constitutes the support service for GoTriple providers together with the Harvesting Management System (HMS) and the helpdesk, which is currently being set-up.

#### Overview:

- The GoTriple handbook for content providers represents a first level of information. It exposes the GoTriple policies, the requirements, the best practices and the support service.
- These policies describe the perimeter and scope of GoTriple. They also specify the kind of data indexed on GoTriple, i.e. publications and research data.
- The requirements for becoming a content provider relate with metadata accessibility (protocols and APIs) and standardization (metadata schemas).
- The handbook provides best practices and recommendations for every metadata field used by GoTriple, in alignment with Dublin Core and OpenAIRE recommendations, as well as with the FAIR principles.
- The final section of the handbook gives a full description of the process and the technical environment through which a content provider can have its content referenced on the GoTriple platform.

## 3. Harvesting process

### 3.1 Harvesting process in the context of the project

In the course of the TRIPLE project, a dedicated platform for data ingestion and curation has been developed. Named SCRE (Semantic Content Retrieval Engine), it processes data by using a pipeline approach: first, it harvests various data sources and then the retrieved information is normalised and enriched.

The technical architecture of SCRE is presented in the D4.4 deliverable while the whole normalisation and enrichment process is described in the [D2.5 deliverable](#).<sup>14</sup> This section provides an overview of the most significant steps of the harvesting, normalisation and enrichment process.

From a conceptual viewpoint, processing in SCRE is managed by two elements: *Sources* and *Flows*. Each Source takes care of data acquisition and processing from a single point of origin (e.g. an OAI-PMH Endpoint or a files archive).

Flows on the other hand take care of the publication of data in the GoTriple index: in particular, flows are used to aggregate data of a specific “provider” presented in the GoTriple front-end. This organisation simplifies data management, as very often providers publish

---

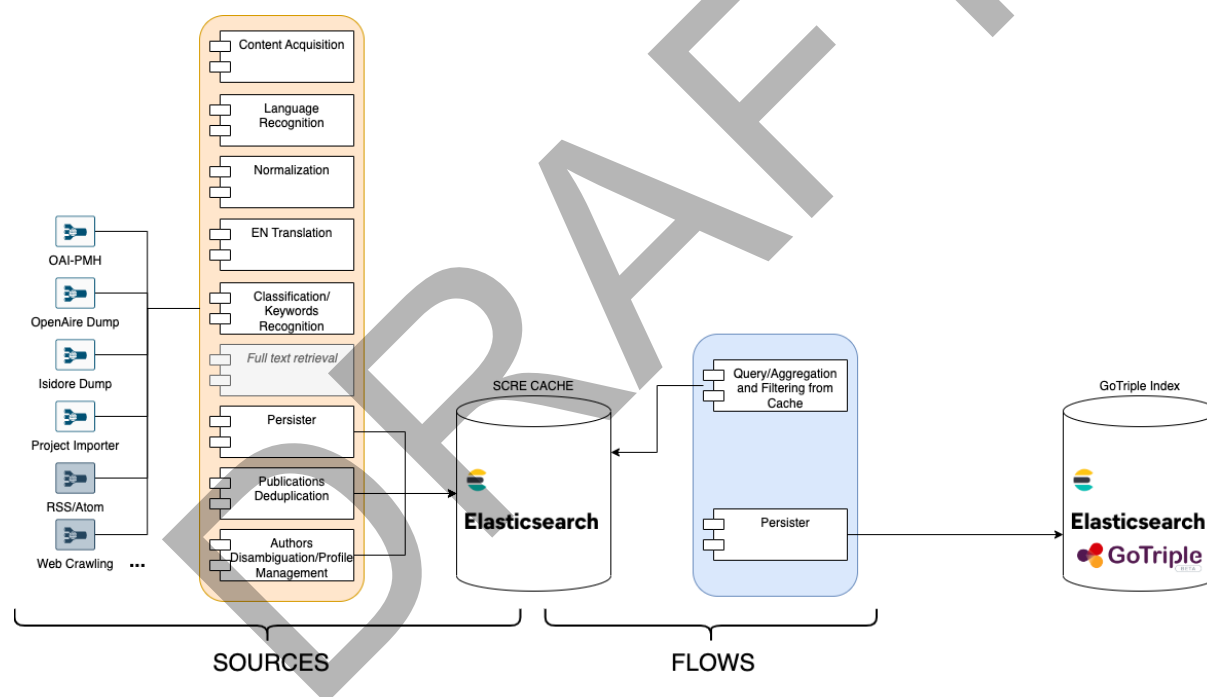
<sup>14</sup> <https://zenodo.org/record/7359654#.Y4YroezMI40>.

data through multiple endpoints: for example data from Biblioteka Nauki are taken from 1,028 sources/endpoints, EKT from 56 and DOAJ from 21.

From a technical viewpoint, on the other hand, there are three main components of the SCRE platform:

- **Connectors:** the components which retrieve metadata about publications and projects from specific data sources
- **Processors,** which curate or enrich the original metadata, according to the logic described herein
- **Persisters,** which finally save the enriched metadata in the platform indexes.

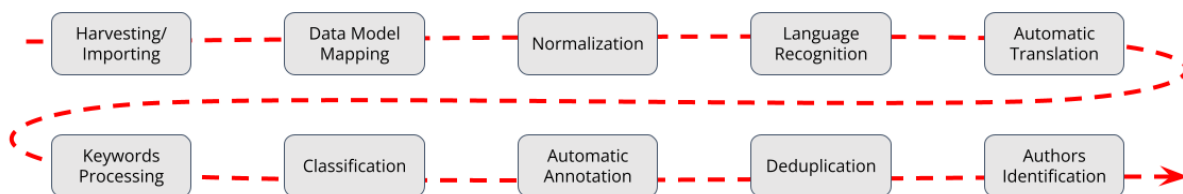
Figure 1 shows a representation of the SCRE technical architecture.



**Figure 1 - SCRE software architecture**

For the purpose of this deliverable, we describe the logic behind SCRE sources, in a “data flow” fashion. Each source is processed in steps: the periodical retrieval of metadata of publications or projects is done by a connector; then, for each metadata record, its normalisation and enrichment are obtained by a series of processors; finally, the result is stored, via a persister, in the platform indexes, implemented with the Elasticsearch search engine.

SCRE is able to process two kinds of GoTriple metadata: publications and projects. The data flow of their processing is presented in Figure 2 and Figure 3 respectively.



**Figure 2 - Publications data flow**



**Figure 3 - Projects data flow**

As far as the publications are concerned, their metadata are acquired from various sources by dedicated SCRE connectors.

In the course of TRIPLE, the following three kinds of these specialised components have been implemented:

- **OAI-PMH connector.** It is the most widely used connector, since a significant number of aggregators and data providers offer a dedicated endpoint to allow third parties to harvest their repositories via this standard protocol. Information is returned in XML format by using several data models: the most common is Dublin Core (DC), whose mapping has been implemented in GoTriple together with the Europeana Data Model (EDM). Data is periodically retrieved by querying an OAI\_PMH endpoint.
- **OpenAIRE connector.** OpenAIRE provides GoTriple with data dumps, consisting of a collection of files in their specific JSON format, aggregated and compressed into a single .zip file, corresponding to a selection of their SSH-related publications. This connector scans the local file system where an OpenAIRE dump is located and imports all of its files.
- **Isidore connector.** As in the case of OpenAIRE, the Isidore publications metadata are made available as files in the local file system of SCRE, formatted in XML and organised in a hierarchy of directories.

After retrieval, publications metadata are normalised, a process which proved to be anything but straightforward. Several problems have been faced while new data sources were added

to the platform: data quality issues of the original metadata, mismatch with the GoTriple data model, the use of free textual strings for structural attributes, and custom extensions of standards by data providers, are just some of the difficulties that have been met.

The data normalisation components aim at producing a consistent, and possibly improved, output by processing the original metadata acquired by the connectors. The metadata elements that have been decided to be subject to normalisation in GoTriple are:

- publication date (date\_published)
- the language of the document (in\_language) and the “lang” attribute of titles (headline), abstracts and keywords
- keywords
- the document type (additional\_type)
- the licence
- the access rights (conditions\_of\_access).

These metadata were chosen for normalization because they are used for refining the results of a search, by using “faceted filtering”. Therefore it was paramount to present consistent representations and to limit their possible values by using controlled vocabularies.

Also, authors’ names are normalised to take into account multiple spellings. For example, the same Greek author might be indicated as such:

```
<dc:creator xml:lang="en">Kapanidis, Nikolaos</dc:creator>
```

```
<dc:creator xml:lang="el">Καπτανίδης, Νίκος</dc:creator>.
```

A dedicated processor tries to recognise these situations to avoid erroneous duplications of authors.

GoTriple always maintains the original metadata received: normalised values are therefore copied into separate elements of the final GoTriple Publications index on Elasticsearch.

After normalisation, an enrichment process ensues, which includes:

- **Language recognition and translation.** This phase has been introduced since some publications miss the language attribute in their textual descriptions (title/headline and abstract) or provide a wrong value for it. Moreover, in order to guarantee the widest comprehension of GoTriple multilingual data, it has been decided to always provide an English translation for these texts if it is missing from the original metadata. For this purpose, the eTranslation service has been used.

- **Classification and automatic annotation** consists of two independent services that expose APIs that, given a text, return:
  - for classification, one or more “discipline”, that is the [27 MORESS](#) (more to see in [D2.3 Report on Machine Learning](#)) categories that have been selected in the TRIPLE project as representative for the SSH domain
  - for automatic annotation, one or more concepts from the GoTriple Vocabulary, a set of over 3,300 SSH entities was created in the TRIPLE project.

Both vocabularies are multilingual and support the 10 official main languages of GoTriple(IT), Polish (PL), Portuguese (PT), Spanish (ES), Ukrainian (UK).

- **Identification of duplicate publications** was necessary as GoTriple integrates large aggregators, which quite often harvest documents from the same SSH sources. A specific heuristic has been introduced to identify whether a newly ingested publication is a duplicate of one already present in the GoTriple index. This algorithm uses as parameters several attributes of the publication, including the DOI, the title, the year of publication, the number of authors and the publisher.
- **Disambiguation of authors.** Authors found in a publication automatically are added in the Profile index of GoTriple. A specific disambiguation procedure has been introduced to recognise the same authors when they appear in multiple publications. This task proved difficult since a single person might be spelled in a variety of ways, e.g. “Suzanne Dumouchel”, “Dumouchel, Suzanne”, “Dumouchel, S.” and also there might be homonyms among them. The disambiguation procedure is based on a set of rules which take into account, given a publication, the name of every single author and its possible variants (e.g. “Dumouchel, Suzanne” -> “Suzanne Dumouchel”, “Dumouchel, S.”), the year of the publication, the publisher and the keywords of the publication.

Finally, it is worth mentioning the processing of projects’ metadata. There is a general lack of structured metadata concerning scientific projects, and the few existing ones are rarely compliant (e.g. difference in granularity or richness level). The TRIPLE team managed to list a small number of potential European and national sources, but the available information didn’t allow for selection of SSH-related projects, nor for their meaningful integration. It has been decided therefore to import data only from:

- **CORDIS:** data about EU-funded projects, related to the SSH domain, under the FP7, H2020 and Horizon Europe research work programmes. CORDIS project metadata are periodically downloaded as Excel files aggregated and compressed in a zip archive. We only process projects that can be identified as Social Sciences or Humanities through a specific mapping between the European Science Vocabulary (EuroSciVoc)

classification and GoTriple's MORESS categories. Besides that, no particular curation is applied as their description is quite precise and all textual elements are in English.

- OPERAS Crowdfunding projects, implemented as a dedicated channel of the WeMakelt platform (to be implemented in the first quarter of 2023).
- OPERAS' COESO projects: citizen science projects managed through the VERA platform, currently under development and not yet operational.<sup>15</sup>

### 3.2 Harvesting management tool

The tool was designed and implemented as part of Task T2.6, whose aim, as described in the grant agreement, was to “provide the effective management of the harvesting process (by creating) a harvesting management system to enable data providers to declare their content to be harvested and to manage interactively the interactions between data providers, language contact points and the platform technical team”. Thus, the goal was to provide an easy way for content providers to propose datasets to be included in GoTriple.

SCRE content acquisition pipeline, as a parametric and configurable platform, already had a web application to control its operation. Through it, GoTriple administrators can define the rules to acquire data sources and to manage flows, as described in the previous chapter “Harvesting process”. The screenshot in Fig 4. shows an example of the original UI of the SCRE web interface.

The adopted strategy was to open this functionality to data providers and basically to turn this user interface, which wasn't designed for the general public but for administrators, into something more user-friendly.

---

<sup>15</sup> VERA is one of the outcomes of the COESO research project, which aims to develop and sustain citizen science research in the social sciences and humanities. COESO has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020) SwafS-27-2020 – Hands-on citizen science and frugal innovation, under Grant Agreement No.101006325.



## Your sources

ADD SOURCE +

Newest first ▾
Any typology ▾

**Openaire Dump - translation** - January 31, 2022 - 11:48 Preview Edit Delete

Has **10000** elements and has received its last element on **February 15, 2022 - 11:00**

Used in **0** flows

**EKT / Annual Symposium of the Christian Archaeological Society** - December 29, 2021 - 12:12 Preview Edit Delete

<https://e-proceedings.epublishing.ekt.gr/index.php/index/oi>

Has **18** elements and has received its last element on **February 14, 2022 - 23:30**

Used in **1** flows

**Figure 4 - The original UI of the SCRE web application**

Another important point was to guarantee that only the right data providers, that is, those with meaningful data (SSH-related and of satisfactory quality) could be onboarded in GoTriple. Hence, the whole process needed multi-faceted oversight procedures.

First of all, it is necessary to identify the data provider and the kind of datasets they have available.

Then, a verification of the quality of the proposed data sources is necessary, both to control the quality and their connection to the SSH domain.

The Harvesting Management System (HMS) is therefore a derivative that evolved from the more general SCRE web interface for controlling the operation of the platform. Hereinafter we only refer to HMS for the whole functionalities of the SCRE web interface.

Firstly, a more fine-grained role system was implemented, to distinguish the different kinds of users accessing the system. The full set of roles available includes:

**Three roles within the GoTriple Team:**

- **Administrator:** manages all aspects of the platform. Only administrators can accept or reject data providers registration requests (see below)
- **Content editor:** reviews the proposed data sources and decides whether to accept or reject them.

- **Viewer:** has access to the statistics of the sources and flows managed by the platform.

#### Role outside of the GoTriple Team:

- **Source Proponent:** the role for data providers who register to the HMS to propose their data sources.

The other important extension was to implement a coherent workflow for onboarding of data providers. This work was preceded by defining the following “user journeys”, for data providers (DP), on the one hand and GoTriple Content Editors (CE), on the other. In the following section we discuss both user journeys, which constituted the user requirements for the implementation of the service, and the actual implementation of requested features in the HMS by including screenshots of its user interface. The HMS is publicly accessible at the URL <https://pipeline.gotriple.eu>.

#### *User Journey 1 - Registration*

- The DP fills up a web form on a public web page of the HMS to register as a provider. Fig.5 provides the screenshot of the actual form used for this purpose: a data provider must specify a description of their institution and the data proposed for GoTriple.

The screenshot shows a registration form for data providers. At the top left is the 'Triple SCORE' logo. The form contains the following fields: 'First Name', 'Last name', 'Email address' (with a dropdown arrow), 'Institution or publisher', 'Textual presentation', 'URL', 'Motivation to be onboarded in TRIPLE', 'Description of the provided dataset', and 'Details to get a sample of data'. Below these fields is an orange button labeled 'Create new account'. At the bottom of the form, there is another 'Triple SCORE' logo and a small copyright notice: 'Copyright © 2022 GoTriple'.

**Figure 5 - Registration form for Data Providers in the HMS**

- A CE checks the registration requests and approves or rejects them. As shown in the screenshot that follows, a DP is automatically registered in the platform with the “Blocked” status. The access to HMS is granted by the administrator who verifies the request and changes the status to “Active” (Fig. 6).

**First Name \***  
 Jacopo

**Last name \***  
 De Santis

**Email address \***  
 jacopo@versacrum.com  
A valid email address. All emails from the system will be sent to this address. The email address is not made public and will only be used if you wish to receive a new password or wish to receive certain news or notifications by email.

**Password**  
 [ ]

**Confirm password**  
 [ ]

Passwords match:

To change the current user password, enter the new password in both fields.

**Status**  
 Blocked  
 Active

**Roles**  
 Source Proponent  
 Authenticated user  
 Administrator  
 Content editor  
 camel  
 Viewer

[▶ LANGUAGE SETTINGS](#)

[▶ LOCALE SETTINGS](#)

**Institution or publisher \***  
 Ver Sacrum

**Textual presentation \***  
 Ver Sacrum is an online magazine regarding Gothic, in every possible form.

**URL \***  
 https://versacrum.com

**Motivation to be onboarded in TRIPLE \***  
 We have a significant collection of music, books and concert reviews.

**Description of the provided dataset \***  
 We can provide our articles through our OAI-PMH endpoint.

**Details to get a sample of data \***  
 https://www.versacrum.com/vs/category/reclusica

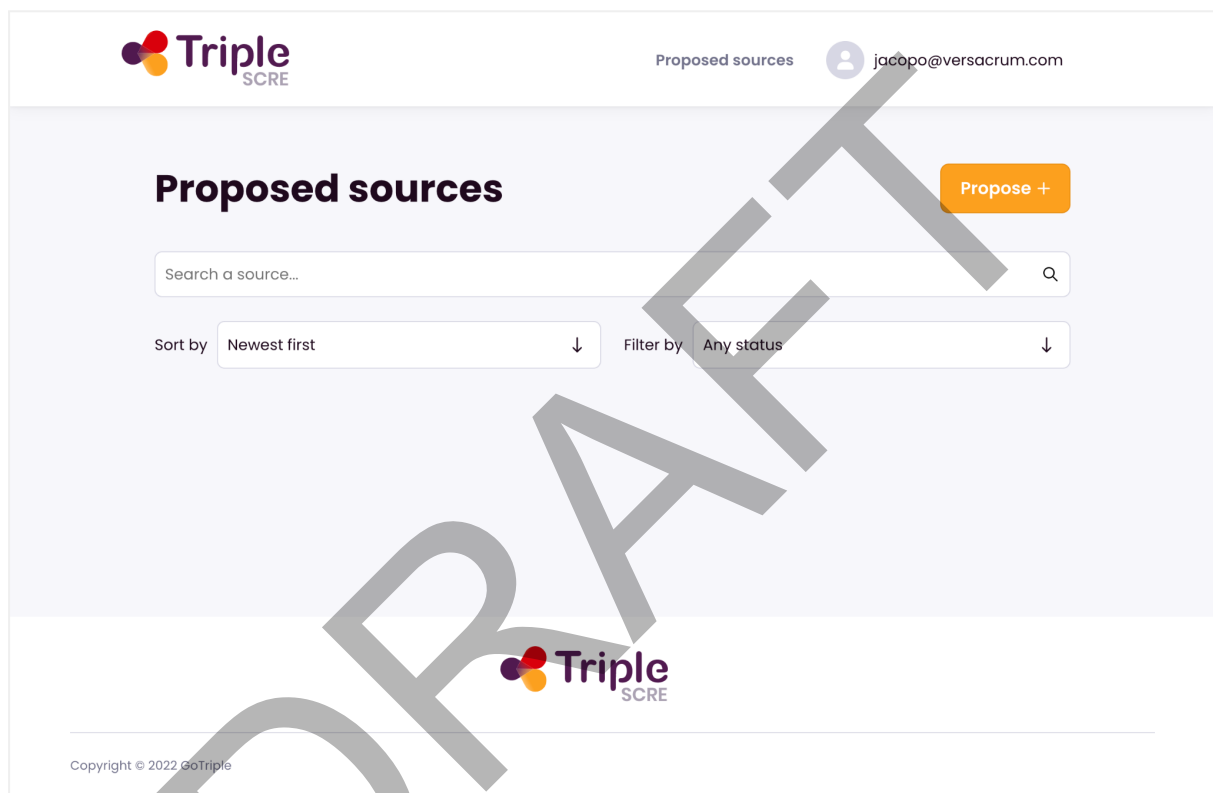
[Save](#) [Cancel account](#)

**Figure 6 - Verification of a Data Provider request for registration**

- Once approved, the DP receives an email notification with the assigned credentials allowing the access to HMS (the login, which is the email specified in the registration phase, and an automatically generated temporary password).



## User Journey 2 - Content source proposal and verification

- Once approved, the DP can access the HMS by connecting at <https://pipeline.gotriple.eu/>, and completing the registration by setting a password, and logging in.
- DP accesses user interface of HMS where they can propose one or more content sources (Fig.7).



**Figure 7 - Data Providers home page in the HMS**

- DP can propose a data source through a wizard-like interface organised in three steps:
  - a. describing the source and specifying the URL and the parameters of the OAI-PMH endpoint, in particular the supported metadata format (either Dublin Core or the Europeana Data Model) (Fig. 8);

 Proposed sources  jacopo@versacrum.com

### Propose a new source

Name


Description

URL address

Metadata prefix

Set

[Next step →](#)



Copyright © 2022 GoTriple

**Figure 8 - Proposing a source by a Data Provider: specifying the OAI-PMH endpoint**

- b. verifying the preview of the fetched results. Should the verification yield problems, DP contacts GoTriple team for support (Fig. 9 & 10);

The screenshot shows the Triple SCRE interface. At the top left is the Triple SCRE logo. At the top right, it says 'Proposed sources' and shows a user profile for 'jacopo@versacrum.com'. The main heading is 'Propose a new source'. Below this is a 'Results Preview' section containing two article cards. The first card is for 'A non-photorealistic rendering method based on Chinese ink and wash painting style for 3D mountain models', published on 2022-11-01 by Ming Yan Jie Wang Yinghua Shen Chaohui Lv. It has tags for 'Non-photorealistic rendering (NPR)', 'Ink and wash painting', and 'Winkle rendering', and a '+2' button. The second card is for 'Analytical studies on medieval lead ingots from Wrocław and Kraków (Poland): a step towards understanding bulk trade of lead from Kraków and Silesia Upland Pb-Zn deposits', published on 2022-11-01 by Beata Miazga Paweł Duma Paweł Cembrzyński Milena Matyszczał Jerzy Piekalski. It has tags for 'Lead', 'Ingot', and 'Trade', and a '+3' button. Both cards have a 'Preview' button and a truncated abstract.

Figure 9 - Proposing a source by a Data Provider: verifying a preview of the fetched results

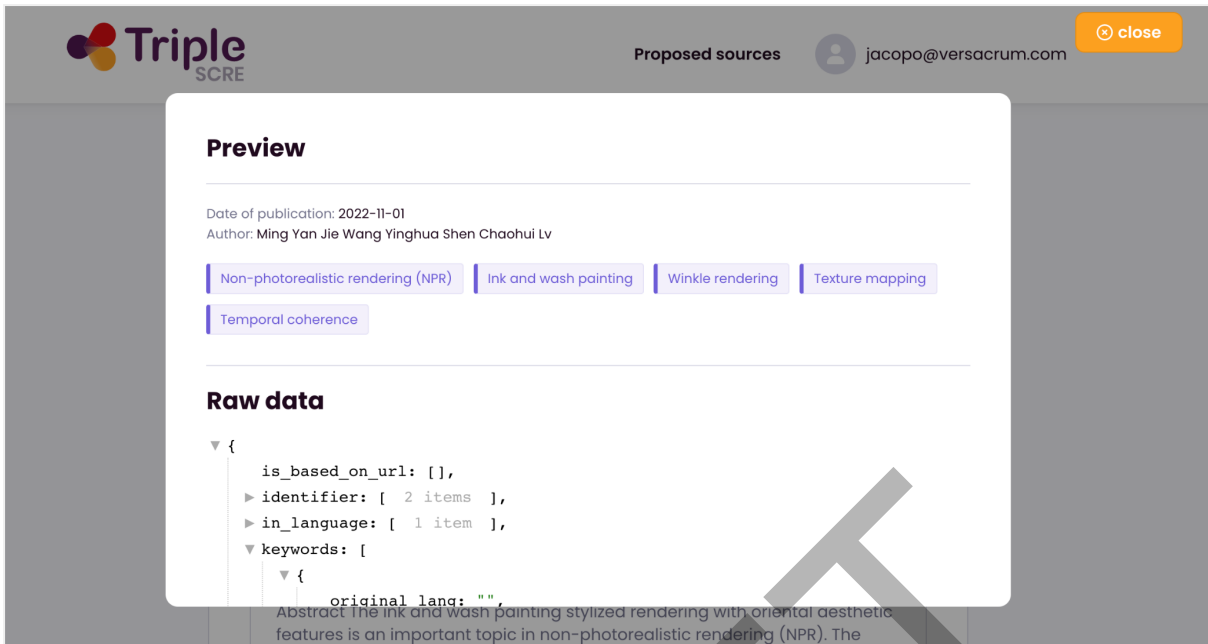


Figure 10 - Proposing a source by a Data Provider: detailed view of a fetched result

c. confirming the proposal of the source (Fig. 11).

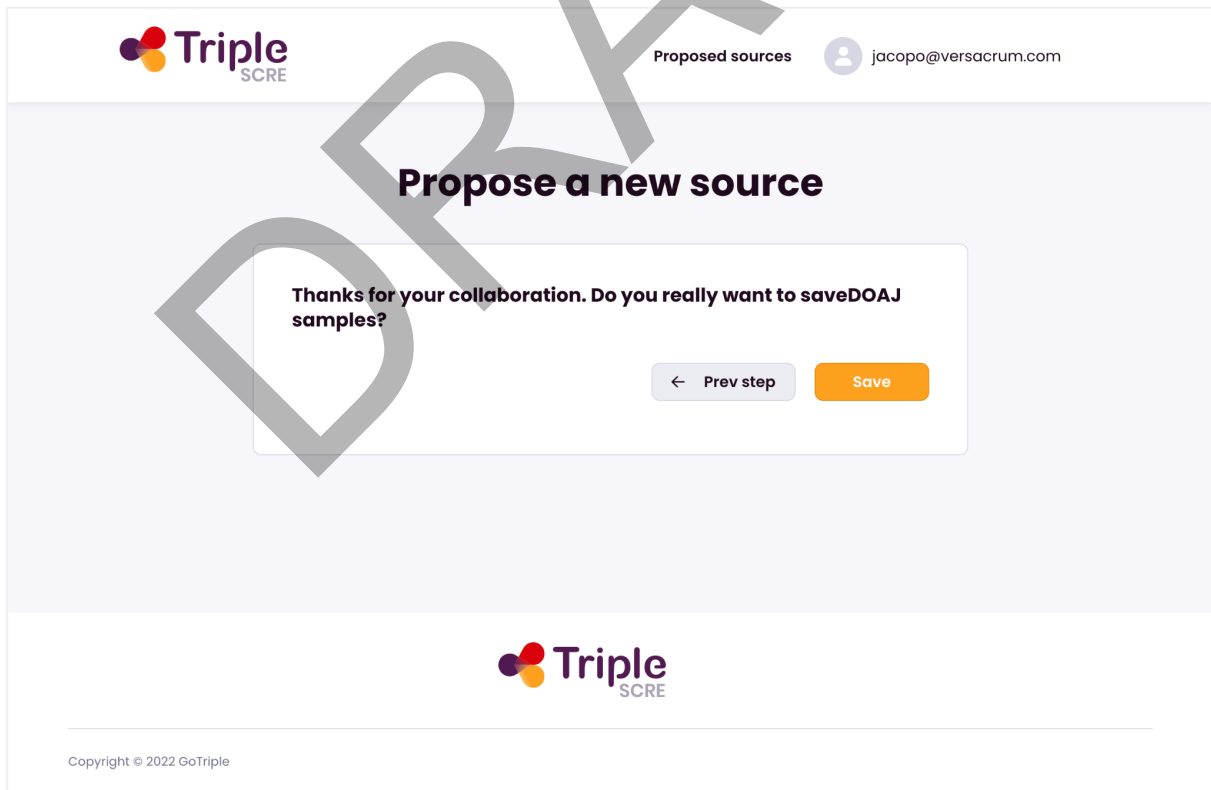
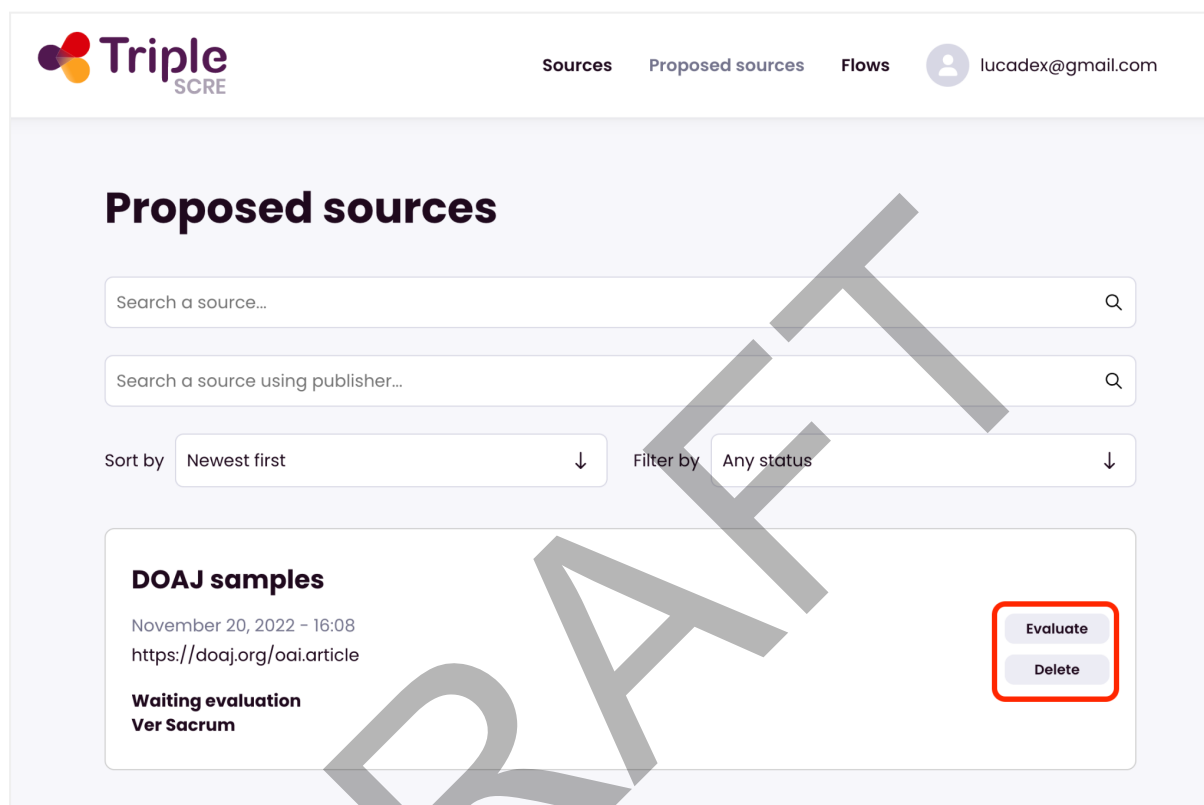


Figure 11 - Proposing a source by a Data Provider: final confirmation step

- A CE controls the proposed content sources and accepts or rejects them (Fig. 12).



**Figure 12 - Content Editors view to accept or discard a proposed source**

- The DP can always see their proposed sources but cannot modify those already approved. In order to do that DP needs to contact the GoTriple team by email for that.

*User Journey 3 - Proposed sources approved and ready to be acquired in GoTriple*

- An approved data source becomes a regular source of the SCRE platform. The CE can check the content retrieved and analyze the statistics to assess its quality.



- If the source proves to be valid the CE either creates a new “Flow” dedicated to it or selects an existing one. With this latter step, the data source is programmed to be ingested, processed and published in the GoTriple platform.

#### User Journey 4 - Assessing the quality of the data retrieved by SCRE

In order to verify the quality of the retrieved content, two features have been implemented in HMS :

- Preview;
- Statistics.

They are both available for sources and flows. The former allows users to view a sample of the content related to that source/flow: for each element it is also possible to access a direct view of the data as it is stored in the platform’s Elasticsearch indexes (those of the SCRE cache for sources and those of the GoTriple front-end in the case of flows). In case of any problems or discrepancies detected, DP contacts CE for support and advice.



**Figure 13 - Source preview**

Statistics on the other hand present a summary view of all retrieved data, again as stored in the SCRE cache, in the case of sources, or in the GoTriple indexes for flows. The two screenshots below show the HMS statistics in these two cases.



Figure 14 - Statistics of a source and of a flow

## Conclusions

### Lessons learned

The data sources selection process was long and laborious but allowed for identification of appropriate sources for GoTriple and efficient construction of both architecture and the acquisition process itself.

The first step of data acquisition was the identification of data sources, aggregators or providers. While the identification of aggregators was a rather straightforward task, it appeared that they did not always provide data in a satisfactory form (due to low standardization, no automated retrieving protocol) or that they faced difficulties with selecting SSH contents in their collections. In these cases, the solution was to retrieve data through a meta-aggregator allowing both automated harvesting and content selection (e.g. the SSH collections of NARCIS are retrieved through OpenAIRE). In the case of providers, on the contrary, the integration of data is more straightforward as the collections are sometimes already SSH-specific and the data is well curated. However, the identification of a significant number of providers on a European scale proved to be more difficult, given the high atomization of the SSH environment. Nevertheless, the process of overcoming these obstacles and challenges resulted in a high coverage of the data aggregated at the European level, and led to an improvement of the providers' support (HMS, Handbook) that will hopefully increase the number of providers engaging with the platform.

Other challenges were related to the type of data considered. While the amount of documents (publications and datasets) harvested by the platform is significant, it has been more challenging to obtain similar results for profiles and projects. Regarding profiles, the project proved that despite the widespread use of authority files and PIDs for authors, they are not always listed in the metadata. This is an area where GoTriple can play a role in the future, for the mutual benefit of the platform and its users. Harvesting the project data proved to be even more challenging. With the exception of CORDIS and some information from the French ANR, availability of standardized and updated information about past and present projects seems to still be a gap in the ecosystem: for example, when information is available, it is sometimes only recorded in a spreadsheet, with uncertain geographic and temporal coverage. The simple fact of highlighting the projects as a searchable object, like the GoTriple platform does, could be a first step to encourage a collective effort to address the issue properly.

Another important challenge concerned metadata quality, which is obviously a major concern for aggregators like GoTriple. Even in the cases where standards are used, and even rather simple ones like Dublin Core, there is still room for interpretation of the standards, and thus for a high variety of the practices. This aspect came fully into light in the GoTriple data acquisition process. The project established various strategies to address the issue at

multiple levels: the set-up of a unique data model, its mapping with other major data models, the normalization steps, and the semantic enrichments. However, it seems necessary to complement these strategies by increasing the interactions between the platform and the data providers to fully address the issues of metadata quality. This is precisely the goal of the HMS and the handbook, and one of the priorities of the GoTriple future developments and enhancements.

## Next steps

The management of data acquisition within GoTriple will rely upon the Data and Tools component of the future GoTriple Committee. In terms of sources selection, the proposition and validation process will be overviewed by the committee on the basis of the existing policies and in compliance with the strategies of the OPERAS RI and its partners in GoTriple's maintenance. In terms of support, besides the HMS and the handbook, the setup of GoTriple as a dedicated service at the level of OPERAS RI will allow to address specific requests for sources addition or harvesting issues. On a more technical level, a major enhancement will concern the links between data and publications. GoTriple already covers both data and publications, however there is often no automated way in the existing metadata to connect them. Ongoing projects in which OPERAS RI and TRIPLE consortium members are involved are already working on these aspects (e.g. with the use of the Scholix framework), and the GoTriple team will follow the developments in order to integrate them into its data acquisition process.

The future evolution of the platform will require a data processing which is more precise, efficient and scalable. These three concepts cover different spaces of improvements.

As far as the *precision* is concerned, as mentioned earlier, we faced significant data quality issues during acquisition, issues that were only partially solved. The data normalisation process in fact proved anything but straight-forward as several problems occurred with new data sources being added to the platform. The lack of quality of the original metadata, mismatch with the GoTriple data model, the use of free textual strings for structural attributes, custom extensions of standards by data providers, are just some of the difficulties that have been met and to which a possible solution has been proposed and implemented. There is room for significant improvement of this process, on the one hand by applying more sophisticated algorithms (e.g. those based on machine learning and natural language processing for the automated annotation phase) and on the other by specifying stricter normalisation rules, especially when recognising the authors' names in publications. The fact that at present we have already acquired a significant amount of data allows us to better recognise all possible difficult cases, which have been in part already solved in the current platform but definitely there is still room for improvements.

The other two concepts, *efficiency* and *scalability*, are strictly related: the current SCRE pipeline, as indicated herein, processes on a daily basis a significant number of data sources.

One in fact must keep in mind that very often providers and aggregators publish their data not through a single API but by different endpoints, which must be queried independently. Just consider that the current six data providers that offer to GoTriple their documents' metadata with the OAI-PMH protocol, correspond to 1,138 single endpoints that are harvested independently. At present the SCRE pipeline is capable of processing between 200,000 and 300,000 documents' metadata per day. While this number seems remarkable per se, one should consider that the acquisition of large providers like BASE or Europeana, or the re-import of those already harvested when an improvement in the data enrichment has been deployed, is still a too lengthy process, that can last weeks, if not months.

Improvements in this area must face both the refactoring of the current SCRE pipeline code and its associated services (e.g. the Classification and Annotation services), but also an expansion of the current processing infrastructure where the platform runs, which is composed of a single virtual server. In particular the possibility to horizontally scale the processing of data, by adding more parallel virtual servers to share the computing tasks, would significantly increase the overall performance of the system.

Finally, one important area of development for a future evolution of the platform is to consider GoTriple as a new and independent dataset. All the actions that can improve the exploitation of its data should be pursued and encouraged. For example, one of the current on-going experimentations the GoTriple team is doing in these last months of the project regards the definition of a formal GoTriple ontology and an extension of the platform search APIs to return data formatted in JSON-LD format, according to this ontology and to the linked data principles. This is a first attempt to integrate GoTriple with the Linked Open Data cloud and to encourage a fuller and wider reuse of its data.

To conclude, the data retrieval process developed in the TRIPLE project has been developed and tested against the numerous obstacles on many levels, ranging from access points and data selection, to data models and normalisation. This long journey proved bumpy and adventurous but also successful and efficient, as the core procedures were established and put at work with a clear vision of the future improvements and development.