Thinking about the Architecture of the Discovery Platform of the TRIPLE Project

Mélanie Bunel, Laurent Capelli, Jean-Luc Minel, Stéphane Pouyllau

Thinking about the Architecture of the Discovery Platform of the TRIPLE Project

Mélanie Bunel[1], Laurent Capelli[1][0000−0002−1873−3857], Jean-Luc Minel[2][0000−0001−6253−6722], and Stéphane Pouyllau[1][0000−0002−9619−1002]

[1] TGIR Huma-Num UMS 3598,
CNRS, Aix-Marseille Université, Campus Cordorcet, Paris, France
[2] Université Paris Nanterre, Laboratoire Modyco UMR 7114,
Nanterre, France

Abstract. This paper proposes three scenarios developed on the basis of the study of the TRIPLE project proposal, the results of users needs study conducted by a TRIPLE project partner (Abertay University) and the evolution of existing aggregation platforms (ISIDORE, OpenAire, Mosa, Narcis, etc.) which give access to scholar's productions. It aims to outline selection criteria, in terms of scientific and technical information and software architecture.

Keywords: Triple Project, information retrieval, linked open data, semantic enrichment

1 Preamble

The TRIPLE project is an H2020 project funded by the European Union developed by a consortium of 19 partners[3]. TRIPLE aims to develops a full multilingual and multicultural solution for the appropriation of SSH resources [1]. In order to clarify our proposals we first define some terms used in this paper.
BUILD processing chain abbreviated in BUILD chain: a BUILD chain is a chain of software components developed by an operator to manage an aggregation platform.
Operator: an Operator is an organization which manages aggregation platform. He can be public (Huma-Num, Open Edition), private (Google, Academia, ResearchGate) or associative (OpenAIRE, Allen Institute).
Aggregation platform: an aggregation platform is a platform which allows users to access publications, projects or user profiles. There are currently many platforms that cover national or international areas. These include Google Scholar, Isidore, Narcis, OpenAIRE, MOSA, ResearchGate, Academia, SemanticScholar and many platforms offered by university consortia or scientific publishers (Open Edition, Springer, Sage, Cairn, Persée).
Our propositions are intended to be consistent with the objectives stated in Triple's proposal [1], in particular this one:
" The TRIPLE solution will be constituted by a full set of services which altogether, will provide to researchers a first-class experience of SSH resources discovery through a variety of complementary means: The main one, which ensures the feasibility of the project, is the already existing search engine for Social Sciences and Humanities data called ISIDORE, developed by the TGIR Huma-Num (CNRS) and will be used to discover data, but also projects and profiles. Projects and profiles will be added to TRIPLE by the researchers themselves either by completing their profiles or by importing information from ORCID, for instance. " (p. 9)
We also seeks to address user needs expressed in D3.1 (Report on User Needs) [2]. The excerpt from the deliverable copied below illustrates this type of need:

[3] The TRIPLE project (https://www.gotriple.eu/), which is financed under the Horizon 2020 framework (https://cordis.europa.eu/project/id/863420), under Grant Agreement No. 863420, with approx. 5.6 million Euros for a duration of 42 months (2019-2023)

Fig. 1. Excerpt from delivery D3.1 [2]

As these two excerpts show, future users insist on the one hand on needs that are not met by current platforms "*lacking tool to make connections and to develop broad overview of the research topic*" and on the other hand on the abundance of existing platforms *"too many software platforms"*. The analysis of the user needs expressed in deliverable D3.1 [2] and the feedback from the ISIDORE project [3] in which three of the authors of this article had been involved lead us to formulate the following proposals.

1.1 Proposals

Building a Light and Labile Thesaurus The TRIPLE proposal [1] proposes *"TRIPLE will produce Multilingual Thesauri in 10 languages. These will be published as Linked Open Data (LOD) datasets in RDF using Simple Knowledge Organisation System29 (SKOS), a W3C recommendation for the representation of Semantic Web controlled vocabularies. They will also be published as TermBase eXchange (TBX) datasets"* (p. 36). and point out *"further enrichments and more detailed controlled vocabularies will add other levels of description. In this task, existing classifications and controlled vocabularies will be gathered and compared. An evaluation will occur, able to define the possible mapping between those at the level of TRIPLE, using for instance multilinguistic thesauri alignment tools (e.g. Opentheso or BBT) or collaborative tools for ontology creation (semantics.gr). This work will serve as a basis for the creation of new vocabularies required for the*

description in new languages." (p. 45).
Based on the experience of the construction of the ISIDORE platform[3] we suggest to build a lighter thesaurus based on existing national thesauri used in ISIDORE plateform[5] (LCSH, Rameau, BNE, etc. - see `https://isidore.science/vocabularies`), this thesaurus will align concepts in the 9 languages indicated in the TRIPLE proposal. The number of concepts in the thesaurus will be around 3000. It will avoid loss of information during the data enrichment phase and largely solves the problem of cultural non-interoperability of concepts in the languages processed by TRIPLE. This thesaurus must be labile, i.e. it will be regularly updated by a team of data librarian (or information manager) who will monitor the emergence of new topics. It has to be stressed that an important hurdle has to be overcome: the alignment work of 3000 concepts in 9 languages is still a time-consuming task.

Building a TRIPLE Data Model The proposal proposes *"Metadata records produced by TRIPLE will be published using the following standard vocabularies: Component MetaData Infrastructure, Dublin Core Metadata Element Set and DCMI Metadata Terms27."* (p. 36). Considering that these description languages lack semantic precision we suggest to specify a TRIPLE data model based on an ontology[4] [4] that will be a recommendation for scientific information aggregation platforms (such as ISIDORE, OpenAire, NARCIS, etc.) which will have to deliver their metadata packets (see below) according to this model [4] which is based on a subset of the schema.org ontology (https://schema.org/). The proposed ontology [4] (cf. fig 2) relies on 3 main classes (CreativeWork, Person and Project) and 27 properties.This ontology aims at specifying as completely as possible the semantics of the different fields that describe resources delivered by BUILD chains but does not presume in any way how these resources will be implemented.It will be enriched as needed, while taking care to use subsets of existing ontologies in order to avoid semantic idiosyncrasies.

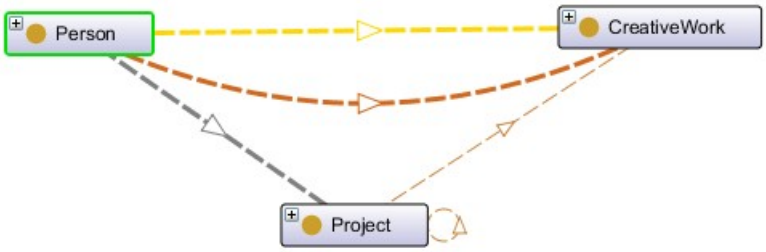

Fig. 2. Main classes and properties of the ontology of TRIPLE data model

It will be up to the operators to transform the harvested metadata, which are generally structured with the DCTERMS vocabulary, into metadata conforming to the TRIPLE model. Deliverable D2.2 [1] (p.45) led by Open Edition partner is in charge to training actions to help providers and operators to bring their data model into conformity with the TRIPLE data model.

[4] An ontology is a formalised set of concepts relevant to a particular area of interest, representing rich and complex knowledge about things, groups of things, and relations between things, as well as a set of constraints about the usage of its terms (cf. https://www.w3.org/OWL/).

Modifying the TRIPLE Processing Pipe Line The proposal describes a pipe line [1] (p.8) composed of software components which harvest and exploit metadata provided by the data providers offering OAI-PMH or SiteMap protocols. It is assumed that TRIPLE pipe line enriches, aligns and indexes these harvested metadata. Taking account it exits already a lot of platforms harvesting metadata we suggest an important change.TRIPLE pipe line will not harvest metadata but it will exploit metadata provided by processing chains, called BUILD-X, managed by operators which already harvest metadata.
A BUILD-X processing chain produces metadata, which will have to be adapted to TRIPLE data model, in json, xml or .tar formats. The BUILD-X chain will deposit packets of metadata on a delivery platform at regular periods to be defined in negotiation with platform operators. This delivery platform is a communication interface between the BUILD-X chains and the TRIPLE pipeline. On this delivery platform, the metadata packets are exchanged on the principle of the push-pull model. In this process, the interaction between a supplier (a BUILD-X chain) and the event channel is push, and between a consumer (TRIPLE) and the event channel is pull. A supplier generates an event and passes it to the event channel. The channel does not transfer the event to registered consumers until consumers pull for the event data. In this process, both suppliers and consumers actively initiate the interaction with the channel and the event channel acts as a queue component. If suppliers and consumers operate at different rates, the channel may need an unbounded queue to store the events. If this is not practical, the channel may need to implement some dropping mechanism or blocking when the queue is full [6]. On this delivery platform, TRIPLE services are capable of:
- Retrieving the resources from the BUILD-X chains and extract them with a software components “Extractor”;
- Checking the conformity of the metadata with the TRIPLE data model;
- Preparing to ingest metadata into the TRIPLE storage system;
- Sending a signal (FLAG-TRIPLE) on the event channel to indicate that the resources has been loaded.

In the long run, TRIPLE will be able to exploit the metadata packets provided by the processing chains named BUILD-O (for OpenAire), BUILD-N (for NARCIS), and BUILD-X (for any platform compatible with the recommendations given by TRIPLE. As not all platforms provide enriched metadata, TRIPLE will develop components that rely on the ISIDORE onDemand[5]to carry out these specific tasks (`https://rd.isidore.science/ondemand/en/`). The interest of this choice is to avoid developing a platform with a specific processing chain when there are already many of them and when these platforms have strongly capitalized their experience. It delegates to these platforms the harvesting and a first step of structuring the harvested metadata. TRIPLE focuses on the enrichment, the linking of these structured metadata and the innovative services that no other platform is able to provide. In return, TRIPLE (via a REST API) offers all platforms that respect the principles of EOSC, access to the metadata aggregated and enriched by TRIPLE.

The proposed architecture makes it possible to position TRIPLE in a dimension complementing existing platforms and defining a pan-European service layer that is currently lacking.

1.2 Time Schedule

We suggest two realization periods, a first one called Build-I, a second one called ALLBUILD

BUILD-I Period In this step, TRIPLE relies on the BUILD-I chain (currently operational in ISIDORE) in order to be able to provide test sets to other partners who need to test innovative services (WP5). Eventually, a test set will be handcrafted to create a "test bench" to design proofs of concept for innovative services; The TRIPLE thesaurus in 9 languages (in coordination with

[5] https://co-shs.ca/fr/nouvelles/entrevue-avec-huma-num-sur-isidore-a-la-demande/

task 2.4 [1] led by The National Documentation Centre (EKT/NHRF) partner will be built. New software components specific to TRIPLE will be built to pull up packets provided by the BUILD-I chain. Metadata extracted from packets will be enriched relying on the TRIPLE thesaurus and the ISIDORE OnDemand APIs (in coordination with the TGIR Huma-Num). Then they will be stored into the TRIPLE database. It is important to note that no assumptions about the type of metadata is made. The enrichment is made from the title, the abstract, the keywords of the resource, whether this resource is a Document (Publication or Data Set), a Project, a Person. The TRIPLE REST API will be specificied and open to EOSC compatible platforms.

ALLBUILD Period This period will be devoted on the one hand to negotiations with the operators of existing platforms (OpenAire, Narcis, ResearchGate, Academia, research funding agencies, etc.) in order they will deliver packets of metadata compliant with the TRIPLE model. On the other hand, to work closely with WP5 led by NeT7 partner to verify that the TRIPLE API corresponds with the needs of the innovative services.

Based on the above requirements, three scenarios are possible, two of which, S1 and S2, meet all requirements. The third scenario S3 is evoked in order to consider all the possibilities.

The preferred scenario, promoted by the authors is scenario S2.

1.3 Scenarios

Scenario S1 In this scenario, the TRIPLE thesaurus is used by a BUILD-X chain of an operator to enrich and categorize the metadata managed by this operator. Consequently an operator has to modify the BUILD-X chain he is managing. A BUILD-X chain must proposes three types of resources (documents, projects, profiles) that are linked together by the BUILD-X chain. All resources are compliant with the data model proposed by TRIPLE. These resources are delivered in the form of packets, to be integrated into a database. TRIPLE pipe line builds an index of these metadata and an API that is exploited by the innovative services developed by the WP5. It is important to point out that at the end of the TRIPLE project, TRIPLE-specific software components are very light. The maintenance service is reduced to the administration of a database and the TRIPLE API, the updating of the TRIPLE thesaurus and the contractualisation between TRIPLE and the partners.

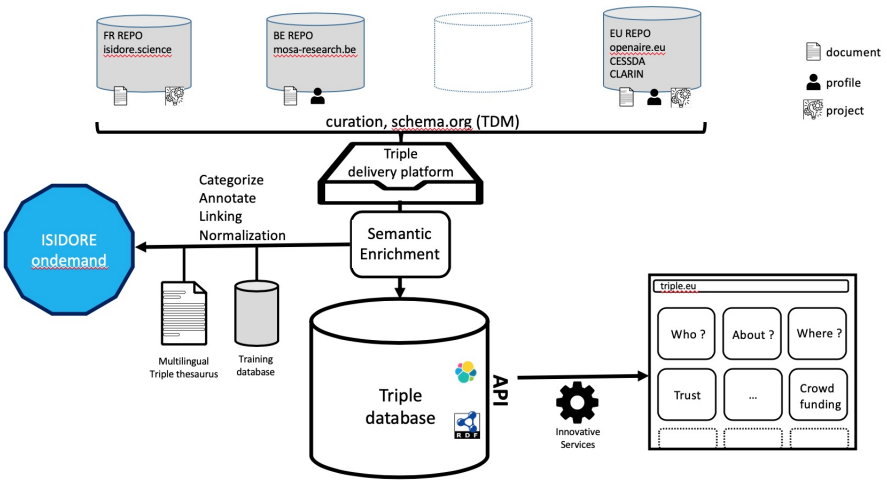

Fig. 3. Scenario S2, Diagram by Laurent Capelli

Scenario S2 Platforms that have contracted with TRIPLE harvest metadata and process them with their BUILD-X chain. An operator does not need to modify its BUILD-X chain. Metadata are delivered periodically by operator in the form of packets on the delivery platform. A packet contains structured and annotated metadata. There are three types of resources :packets of documents (Publications and data set), packets of projects, and packets of researchers which correspond to the four types of metadata managed by TRIPLE. A BUILD-X chain can deliver only a subset of a packet type. It is important to point out that the BUILD-I chain implemented by ISIDORE is not modified and the TRIPLE thesaurus (aligned 9 languages) is not used by BUILD-I ; TRIPLE downloads packets delivered on the delivery platform by the different BUILD-X chains. TRIPLE develops a component that exploits the ISIDORE OnDemand API in order to annotate and categorize resources of all types delivered by the BUILD-X chains. The annotation process is performed using the TRIPLE thesaurus. All the resources (annotated and categorized) and aligned in 9 languages. TRIPLE builds a database and an RDF warehouse that aggregates all the metadata delivered by platforms. TRIPLE indexes the database via ElasticSearch. The database is accessible via a REST API which is exploited by the innovative services developed by the WP5. Operators can possibly, if they are interested, recover the fruit of the enrichment.The RDF warehouse is accessible in the form of a dump.

In this scenario, the BUILD-I base is used as a test bench. The targeted trajectory is on the one hand to propose TRIPLE software components capable of integrating metadata from different BUILD-X chains. On the other hand, to propose to operators methodological and technical tools that allow them to make their platforms fully compatible with TRIPLE. Discussions will be held with the various operators of the BUILD-X chains (TGIR Huma-Num, OpenAIre, Narcis, Mosa, etc.) so that in the medium term they can deliver all types of resources (Documents, Projects, People). It is important to point out that at the end of the TRIPLE project, there are software components specific to TRIPLE. TRIPLE's innovative services via their interfaces are exploited by the project partners and offered to the EOSC community by OPERAS. The maintenance service requires the maintenance of the software components, the administration of a database, the updating of the TRIPLE thesaurus and the contractualisation between TRIPLE and the partners.

The two figures illustrate the interaction between the platforms and Triple's pipe line. Figure 3 presents the general principles of the interactions while figure 4 emphasizes the functioning of the BUILD-X chains.

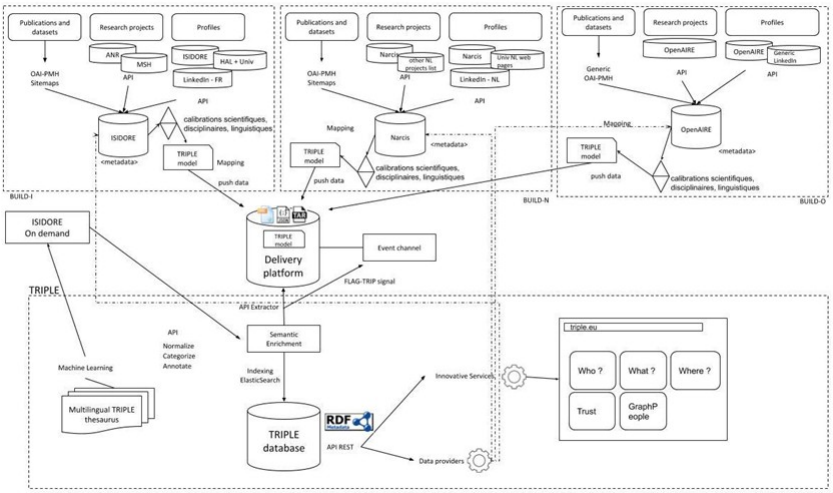

Fig. 4. Scenario S2, Diagram by Stéphane Pouyllau

Scenario S3 This scenario involves the construction of a set of software components, which we call the BUILD-T chain. The construction of this chain can be carried out according to two options. In the option 1, TRIPLE acquires the existing software components in ISIDORE which implies to engage several negotiations. First, with Huma-Num for the ISIDORE technology to be proposed to TRIPLE.Second with ANTIDOT company concerning the license of the proprietary components (AIS/AFS).Third, developing missing software components in ISIDORE.

In the option 2, TRIPLE develops the software components based on the existing ISIDORE chain and fully develops the BUILD-T chain. This one should be based on Open Source software bricks. To our knowledge, there are no robust open source software bricks that can offer the required functionalities for the harvesting perimeter under consideration. In both options, this scenario requires setting up a large team of developers to develop the software components currently missing in ISIDORE to harvest projects and people.It also requires a technical infrastructure that will host the BUILD-T chain.

1.4 Scenario Synthesis

Scenario 1

Strength
- Few specific software developments outside the DBMS and API
- Harmonization of the annotations made from the TRIPLE thesaurus by the different BUILD-X chains.
- Development of innovative services in coordination with WP5
- Sustainability of the developed services provided

Weaknesses
- Strong dependency on operators who need to modify their BUILD-X chain to enrich their metadata with the TRIPLE thesaurus.
- High upgrade requirement for BUILD-X chains

Scenario 2

Strength
- Ensuring the harmonization of the annotations made from the TRIPLE thesaurus
- Relative independence from operators
- Development of innovative services in coordination with WP5
- Strong differentiation from other existing platforms
- Low upgrade requirement for BUILD-X chains

Weaknesses
- Specific software developments in addition to those required for DBMS and API management
- Difficulty to link resources (Documents, Projects, Experts, Researchers) processed by the operators by algorithms specific to each platform.
- Non-guaranteed sustainability of the services developed at the end of the TRIPLE project

Scenario 3

Strength
- Ensuring the harmonization of the annotations made from the TRIPLE thesaurus
- Complete independence from operators

Weaknesses
- Major software developments without the assurance of being able to rely on open-source software components

- Not very compatible with TRIPLE's time schedule
- Non-guaranteed sustainability of the services developed at the end of the TRIPLE project
- Development of a nth access portal little different from existing portals

References

1. Transforming Research through Innovative Practices for Linked interdisciplinary Exploration TRIPLE. The European discovery platform dedicated to SSH resources. Proposal number: 863420 (2018).
2. Forbes, S, De Paoli, S., Błaszczyńska, M., Maciej M.: Deliverable D3.1, Report on user needs. TRIPLE Project. (2020) DOI=10.5281/zenodo.3925022.
3. Pouyllau, S., Minel, J.L, Kilouchi, S., Capelli, L. : Bilan 2011 de la plateforme ISIDORE et perspectives 2012-2015. Comité de pilotage du TGE Adonis. 1-23 (2012). `https://isidore.science/document/10670/1.bqexsj`
4. Bunel, M., Capelli, L., Minel, J.L, Pouyllau, S. : An ontology for the TRIPLE Data Model (2020). `https://gitlab.huma-num.fr/triple/model`
5. David, S., Minel, J.L, Pouyllau, S., : Documenting some Uses of the Isidore Platform (2011). `https://isidore.science/document/10670/1.lbc7dv`
6. Hewitt, C. : Viewing Control Structures as Patterns of Passing Messages, Artfical Intelligence 8(3), 323-326, (1977)