# Building an Ontology to Specify the TRIPLE Data Model Proposal for a Beta Version

Mélanie Bunel,[1][0000−0002−7314−3892] 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] Modyco UMR 7114, Université Paris Nanterre, CNRS, Nanterre, France

**Abstract.** This article describes the conceptual and pragmatic choices that led to the construction of an ontology whose objective is to prepare the complete specification of the TRIPLE data model.

**Keywords:** Triple project, ontology, document data model

## 1 Introduction

The TRIPLE project is a H2020 project funded by the European Union developed by a consortium of 19 partners[3]. TRIPLE aims to develop a full multilingual and multicultural solution for the appropriation of Social Sciences and Humanities resources [9]. In a first research report [5], we proposed three scenarios for the construction of the TRIPLE Project. In a second report [6], we stressed the need to better take into account the scientific context in which new platforms for access to scientific publications are emerging, the most recent of which propose to use AI technologies (SemanticScholar[4], dimensions[5]).

In July 2020, TRIPLE's technical board decided to follow our suggestion by choosing the scenario that proposes to rely on data harvested and structured by a set of existing platforms (OpenAIRE, ISIDORE, NARCIS). This scenario specifies that the data provided by these platforms must comply with a data model that will be proposed to them by the TRIPLE project.

The purpose of the TRIPLE data model is to describe metadata for publications, datasets, projects and researchers' profiles. The proposal proposes *"Metadata records produced by TRIPLE will be published using the following standard vocabularies: Component MetaData Infrastructure, Dublin Core Metadata Element Set and DCMI Metadata Terms."* (p. 36). Considering that these description languages are limited in their semantic description, we suggested to specify a TRIPLE data model based on an ontology. An ontology is an explicit specification of a conceptualization formalised by a 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 [8,3,2]. In order to rigorously specify this model, we propose in a first step to specify this ontology in the formalism of the languages of the semantic web, more precisely the stack of formal languages, RDF, RDFS and OWL [6].

In this third report, we propose two ontologies, each with a different objective. A first ontology, which we will call "full ontology" (FO), aims at specifying the semantics of metadata as rigorously as possible. A second ontology, which we will call 'Light ontology' (LO), aims to get closer to the data models provided by the aggregators (Isidore, OpenAire, narcis, etc.). We will provide in the document some examples that illustrate the consequences of these two modeling options. Concerning the OWL language, we will use OWL2 whose power of expression meets our needs, especially to express unions of Classes to describe the FO and OWL-lite for the LO .

There is a great deal of work on ontologies and on the objectives pursued through their construction [1]. For our part, we assign the following objectives to this ontology:

[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)
[4] https://www.semanticscholar.org/
[5] https://www.dimensions.ai/
[6] https://www.w3.org/OWL/

- capitalising knowledge for preservation and transmission purposes ;
- some standardization of expert knowledge in the field of scientific information, possibly making it possible to standardize the practices of experts and improve the consistency and usability of applications;
- a reference resource for the work of experts, in particular the adaptation work to be carried out by platform operators to adapt their proprietary data model to the TRIPLE data model.

We would like to stress that the specified ontology is not intended to be implemented unchanged, in particular for performance reasons. The data model that will be built for operational use can thus be simplified by removing classes and properties. Nevertheless, we wanted on one hand to be able to check that the constructed ontology is consistent and on the other hand to carry out some proofs of concept by making queries in SPARQL language to verify that it is possible to navigate in the graph of instances. For these reasons, the ontology is built with the Protégé software [7]. Two OWL files describing FO and LO are stored in the Huma-Num gitlab. [8]

## 2 Full Ontology Description

### 2.1 Methodological Choices

We insist on the fact that our objective is not to build a generic ontology of documentary objects but to contribute to the specification of a data model that must pragmatically take into account the availability of metadata provided by the platforms. Consequently, our choices in terms of classes and properties are a trade-off between our willingness to describe them rigorously and the realization that some metadata will not be available. For example, for some literature studies, a scholar might want to indicate that he or she is looking for publications that do not exceed a certain number of words; but this information is rarely provided by existing platforms, so it is not relevant to specify a wordCount property for a SchorlarlyArticle. The construction process is based on several choices.

First, we rely on existing and referenced ontologies to avoid conceptual idiosyncrasies as much as possible. In fact, all the classes come from schema.org [9] as well as almost all the properties. However, we had to create specific subclasses for business needs.

Secondly, for a certain type of data, it is possible to hesitate between creating Class and ObjectProperty or using a DataProperty [10] especially to represent standardized terms or nomenclatures. In general terms, we used the class DefinedTerm proposed by schema.org (pending class) for represent this kind of metadata.

Thirdly, a recurring issue in the design of an ontology is how time is represented, especially events of finite duration. The CIDOC-CRM [10] offers a generic and powerful solution but relatively complex and more oriented object programming than inference language. As far as the ontology of TRIPLE is concerned, main events to be represented concern mainly successive jobs occupied by a same person. To represent these events, we chose to rely on the concept of Role proposed by schema.org which completely meets our needs.

### 2.2 Classes

We identified 27 classes organized in a hierarchy illustrated in Figure 1. Three main classes represent the objects described by the metadata:

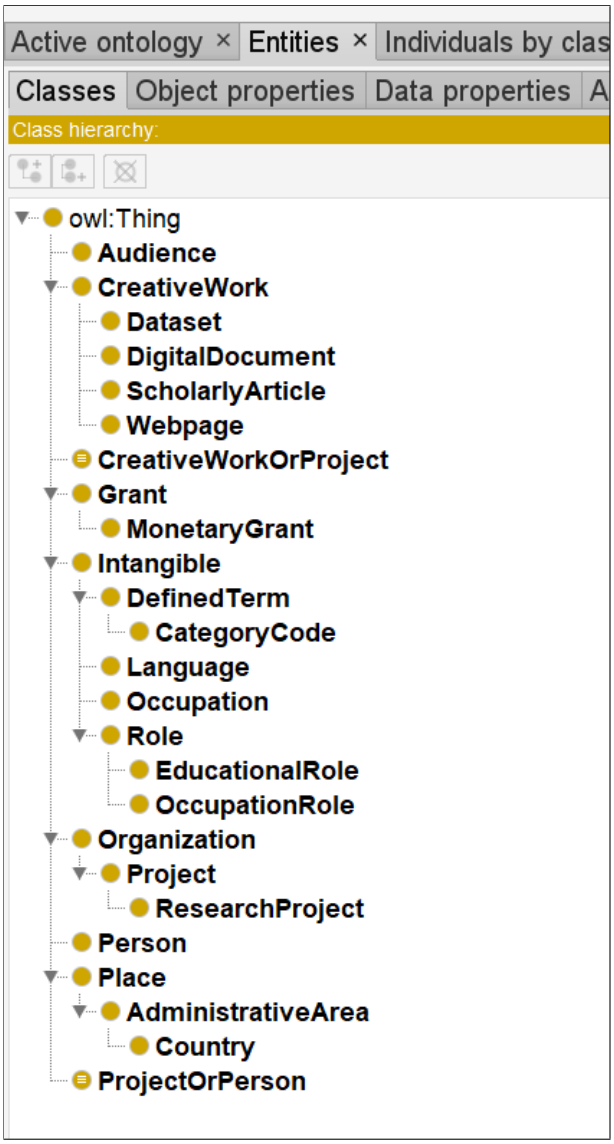
- the CreativeWork class which represents publications and datasets; this class has 4 subclasses:
  - Dataset;
  - DigitalDocument;
  - ScholarlyArticle;
  - WebPage.
- the ResearchProject class (pending class), which represents the class of granted projects.
- the Person class which represents the scholars.

[7] https://protege.stanford.edu/
[8] https://gitlab.huma-num.fr/triple/model
[9] https://schema.org/
[10] http://www.cidoc-crm.org/

Some other schema.org native classes complete this class hierarchy:

- the Audience class characterizes the audience of a publication;
- the CategoryCode class, a subclass of DefinedTerm, is used to describe all nomenclatures or standardized terms;
- the Organization class may characterize several type of organizations like schools, research laboratories, private companies or any other existing types of organizations;
- The Grant class characterizes the funding of a project with one subclass: MonetaryGrant;
- the Language class characterizes the language in which the documentary objects are described;
- the Occupation class characterizes the Person's occupation;
- the Administrative Area class specifies the geographical publication location of some objects;
- the Role Class specifies additional information about a relationship or property.

**Fig. 1.** FO: Class hierarchy



## 2.3 Properties

### 2.3.1 ObjectProperties

We identified 50 ObjectProperties. For each ObjectProperty[11] we specify their Domain, Range, and their cardinalities (Tables 1 to 10). Added comments are generally copied from schema.org. The Figure 2 illustrates how the class CreativeWork is linked to the other classes through the ObjectProperties.

Domain and Ranges are built-in properties define as follow by W3C:

*"Syntactically, rdfs:domain is a built-in property that links a property (some instance of the class rdf:Property) to a class description. An rdfs:domain axiom asserts that the subjects of such property statements must belong to the class extension of the indicated class description. Multiple rdfs:domain axioms are allowed and should be interpreted as a conjunction: these restrict the domain of the property to those individuals that belong to the intersection of the class descriptions. If one would want to say that multiple classes can act as domain, one should use a class description of the owl:unionOf form. "* (https://www.w3.org/TR/owl-ref/#domain-def). In OWL Lite the value of rdfs:domain must be a class identifier, this is one of the reasons why we chose to use OWL2 which allows to express union of classes.

*"For a property one can define (multiple) rdfs:range axioms. Syntactically, rdfs:range is a built-in property that links a property (some instance of the class rdf:Property) to to either a class description or a data range. An rdfs:range axiom asserts that the values of this property must belong to the class extension of the class description or to data values in the specified data range Multiple range restrictions are interpreted as stating that the range of the property is the intersection of all ranges (i.e., the intersection of the class extension of the class descriptions c.q. the intersection of the data ranges). Similar to rdfs:domain, multiple alternative ranges can be specified by using a class description of the owl:unionOf form ".*

Note that in all tables DateTime, Text and URL are DataType. Five ObjectProperties do not have specific Domain because they are used with several classes as Domains [12]: alternateName, identifier, url, stardDate and endDate.

[11] Unless otherwise specified, the name domain is by default schema.org

[12] Nevertheless, for readability reasons we repeat them in the tables below if they are used by the classes indicated in Domain

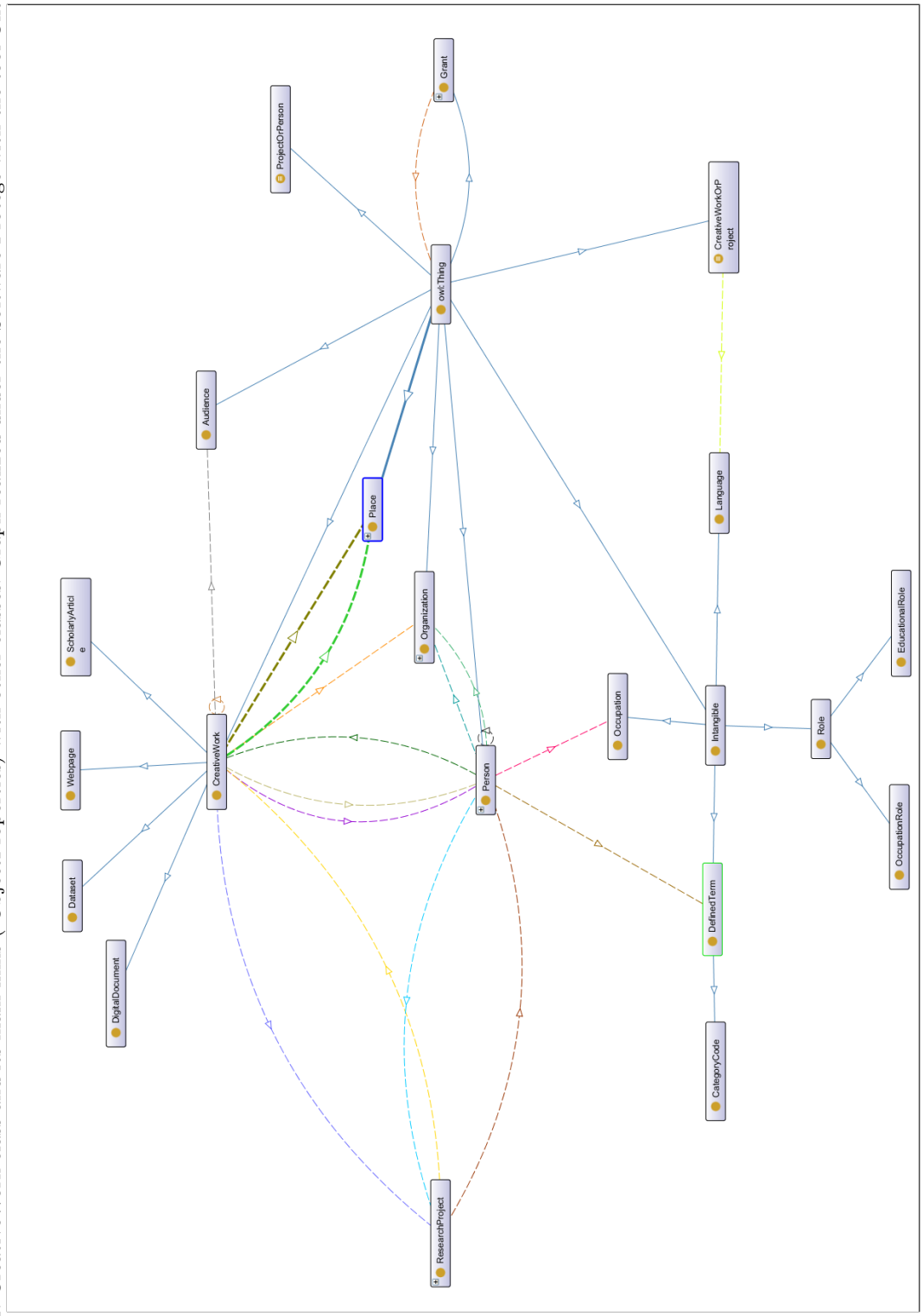



**Fig. 2.** CreativeWork class and its main links ( ObjectProperties) with other classes. Graph realized under the Software Protégé with the tool OntoGraf

**Table 1.** ObjectProperty having CreativeWork as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| abstract | Text | 0:N | An abstract is a short description that summarizes a CreativeWork. |
| about | researchProject | 0:N | The subject matter of the content. |
| audienceType | Audience | 0:N | The target group associated with a given audience (e.g. veterans, car owners, musicians, etc.). |
| additionalType | URL | 1:N | An additional type for the item, typically used for adding more specific types from external vocabularies in microdata syntax. This is a relationship between something and a class that the thing is in. |
| citation | CreativeWork | 0:N | A citation or reference to another creative work, such as another publication, web page, scholarly article, etc. |
| contributor | Person | 0:N | A secondary contributor to the CreativeWork or Event. |
| datePublished | DateTime | 1:N | Date of first broadcast/publication. |
| encodingFormat | Text | 0:N | Media type typically expressed using a MIME format (see IANA site and MDN reference) e.g. application/zip for a SoftwareApplication binary, audio/mpeg for .mp3 etc.). |
| headline | Text | 1:N | Headline of the article. |
| isBasedOnUrl | URL | 0:N | A resource that was used in the creation of this resource. This term can be repeated for multiple sources |
| keywords | Text | 0:N | Keywords or tags used to describe this content. |
| license | CategoryCode | 0:1 | A license document that applies to this content, typically indicated by URL. |
| mentions | Text | 0:N | Indicates that the CreativeWork contains a reference to, but is not necessarily about a concept. |
| publisher | Organization | 0:N | The publisher of the creative work. |
| spatialCoverage | Place | 0:N | The spatialCoverage of a CreativeWork indicates the place(s) which are the focus of the content. |
| temporalCoverage | DateTime | 0:N | The temporalCoverage of a CreativeWork indicates the period that the content applies to, i.e. that it describes, either as a DateTime or as a textual string indicating a time period in ISO 8601 time interval format. |

**Table 2.** ObjectProperty having ResearchProject as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| description | Text | 0:N | A short description that summarizes a Project. |
| funder | Organization | 0:N | A person or organization that supports (sponsors) something through some kind of financial contribution. |
| member | Person | 0:N | A member of an Organization or a ProgramMembership. |
| name | Text | 1:N | The name of the item. |
| parentOrganization | Organization | 0:N | The larger organization that this organization is a subOrganization of, if any. |
| sponsor | Organization | 0:N | A person or organization that supports a thing through a pledge, promise, or financial contribution. In TRIPLE Project case, it is related to the Crowdfunding platform. |

**Table 3.** ObjectProperty having Person as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| alumniOf | Organization | 0:N | An organization that the person is an alumni of. |
| affiliation | Organization | 0:N | An organization that this person is affiliated with. For example, a school/university, a club, or a team. |
| authorOf | CreativeWork | 0:N | A CreativeWork that the Person is author of. This is not a schema property, it has been created as a "triple project" property. |
| hasCredential | Educational(..)Credential | 0:N | A credential awarded to the Person or Organization. |
| familyName | Text | 1:1 | the last name of an Person. This can be used along with givenName instead of the name property. |
| givenName | Text | 1:1 | Given name. In the U.S., the first name of a Person. |
| hasOccupation | Occupation | 1:N | The Person's occupation. |
| jobTitle | Text | 0:N | The job title of the person (for example, Financial Manager). |
| knows | Person | 0:N | The most generic bi-directional social/work relation. |
| knowsAbout | Text | 0:N | Of a Person to indicate a topic that is known about - suggesting possible expertise but not implying it. We do not distinguish skill levels here, or relate this to educational content, events, objectives or JobPosting descriptions. |
| memberOf | ResearchProject | 0:N | An Organization (or ProgramMembership) to which this Person or Organization belongs. |
| nationality | Country | 0:N | Nationality of the person. |

**Table 4.** ObjectProperty having Organization as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| legalName | Text | 1:N | The official name of the organization, e.g. the registered company name. |

**Table 5.** ObjectProperty having Role as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| roleName | Text | 0:N | A role played, performed or filled by a person or organization. |

**Table 6.** ObjectProperty having CategoryCode as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| codeValue | Text | 0:N | A short textual code that uniquely identifies the value. |

### 2.3.2 Restrictions and characteristics

Several same ObjectProperties are used for different Classes. For readability reasons, we prefer to create named classes rather to express restrictions with anonymous classes. This results in "unions of classes" which allows to allocate a same ObjectProperty to 2 classes as following:

- The class called CreativeWorkOrProject (Table 7) is defined as an union between CreativeWork and Project in order to be able to indicate a Domain common to certain properties;
- The class called ProjectOrPerson (Table 8) is defined as an union between Project and Person in order to be able to indicate a Domain common to certain properties;

Some characteristics are also applied:

- the Classes Organization and Person are disjoints;
- the ObjectProperty schema:identifier is Functional;
- the ObjectProperty triple:authorOf is owl:inverseOf schema:author;
- the ObjecProperty schema:knows is Symmetric;
- the ObjecProperty schema:citation is Symmetric;
- all literals conform to RDF 1.1 *Concepts and Abstract Syntax* [13] Precisely, the language is specified by a non-empty language tag as defined by [BCP47].

**Table 7.** ObjectProperty having CreativeWorkOrProject (union of classes) as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| inLanguage | Language | 0:N | The language in which is written the document |
| mainEntityOfPage | URL | 1:N | Indicates a page (or other CreativeWork) for which this thing is the main entity being described. See background notes for details |

**Table 8.** ObjectProperty having ProjectOrPerson (union of classes) as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| description | Text | 0:N | A description of the item. |

[13] http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

Some ObjectProperties are used for more than 2 classes. In that case, the class "Thing" is preferred as Domain (Table 9).

**Table 9.** ObjectProperty having Thing (Generalist Class) as Domain

| Name of ObjectProperty | Range | Cardinality | Comments |
|---|---|---|---|
| alternateName | Text | 0:N | An alias for the item. |
| endDate | DateTime | 0:N | The end date and time of the item |
| identifier | Text | 1:N | The identifier property |
| startDate | DateTime | 0:N | The start date and time of the item |
| url | URL | 0:N | URL of the item (Link to the full text of the document or a webpage). |

### 2.3.3 DataProperties

We identified 1 DataProperty which has Person for Domain (Table 10). For each DataProperty[14] we specify their Domain which is a Class and and a Range which is a literal.

**Table 10.** DataProperty having Person as Domain

| Name of DataProperty | Range | Cardinality | Comments |
|---|---|---|---|
| orcid_id | xsd:string | 0:1 | Authority data on researchers, academics, etc. The ID range has been defined as a subset of the forthcoming ISNI range. |

### 2.4 Checking Ontology Consistency

The HermiT 1.4.3 reasoner from Protégé was used to check the structural ontology consistency, i.e. before ontology was populated. With the exception of a few minor errors in some Range specifications, which were easily corrected, the ontology was validated. Given the small number of restrictions expressed, this quick validation is not surprising.

## 3 Light Ontology Description

To illustrate the difference in modeling between Full Ontology (FO) and Light Ontology (LO), we will take the example of a title of a publication. A publication is an instance of the Creative Work class in both cases, FO and LO.

In the case of FO, the title is associated by the headline property with a string (type xsd: string). The language of the publication and the language of the title is indicated by the inLanguage property. The Range of this property is the Language class. This Language class is associated by the alternateName property to an instance of the CategoryCode class which specifies all possible codes for the languages.

In the LO, the title is associated by the headline property at a rdf:PlainLiteral datatype. A rdf:PlainLiteral lexical form is a string of the form "abc@langTag" where "abc" is an arbitrary (possibly empty) string, and "langTag" is either the empty string or a (not necessarily lowercase) language tag [15]. It should be noted that the W3C recommendation indicates *"To eliminate another source of syntactic redundancy and to retain a large degree of interoperability with applications that do not understand the rdf:PlainLiteral datatype, the form of rdf:PlainLiteral literals in syntaxes for RDF graphs and for SPARQL is the already existing syntax for the corresponding plain literal,*

[14] Unless otherwise specified, the name domain is by default schema.org
[15] https://www.w3.org/TR/rdf-plain-literal/

*not the syntax for a typed literal. Therefore, typed literals with rdf:PlainLiteral as the datatype are considered by this specification to be not valid in syntaxes for RDF graphs or SPARQL"*

Here are some consequences of these two models. First, with FO, it is possible, by using two instances of headline property, to indicate that a title contains a part in one language and a part in another language as for example in the following title, with a latin part and a french one : " « Non obstante quod sunt monachi ». Être moine et étudiant au Moyen Âge" [7]. This is not possible in LO because there is only one tag to indicate the language. Likewise, in the case of LO, it will not be possible to search for titles that are expressed in 2 or more languages whereas it is possible in the case of FO.

Second, in a SPARQL query, in the case of LO, you will have to use the FILTER clause to search for titles in a language [16], which can be very time-consuming to compute.

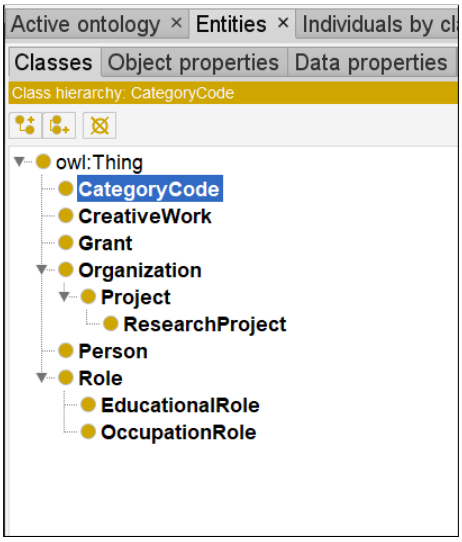
Thirdly, in the case of FO, it is easily possible to know all the language codes used (and accepted) since they are instances of the class CategoryCode. In the case of LO, it will be more difficult and much more time-consuming.

### 3.1 Classes

We identified 10 classes organized in a hierarchy illustrated in Figure 3. As for the FO, three main classes represent the objects described by the metadata:

- the CreativeWork class which represents publications and datasets;
- the ResearchProject class (pending class), which represents the class of granted projects;
- the Person class which represents the scholars.

**Fig. 3.** LO : Class hierarchy




[16] such as FILTER(lang(?title) = 'en')

### 3.2 Properties

#### 3.2.1 ObjectProperties

The LO has the same ObjectProperties that FO. However, some properties, considered as very difficult to obtain and therefore becomes optional, are not included in the LO like schema:audience, schema:nationality, schema:encodingFormat, schema:hasCredential and schema:contributor. The big difference is that there are no constraints on the Range of these properties, which are mainly literal or DataType objects.

#### 3.2.2 DataProperties

The LO has the same DataProperties that FO.

### 3.3 Ontology Graph

To convey the structural relationships in a simple and more readable format, we have adopted a graph-based model to represent the Light Ontology. For the sake of clarity, we formalized the ontology into a global graph with all the classes and their properties (Figure 4, p 12). For educational purposes, this graph contains "orange boxes" to explain to which metadata fields each property related to. Also, the "grey boxes" refer to the information which will be added to the 3 objects during the enrichment process (WP4) after the collection of raw data provided by the aggregators. The enrichment process will consist in:

- Categorization with a field dc:subject referring to SSH disciplines (MORESS Project);
- Semantic annotation with the property schema:knowsAbout.

## 4 Conclusion

It exists a lot of metadata standards used to describe the objects that TRIPLE targets to feed its database. This heterogeneity increases the difficulty to obtain a satisfying level of quality of metadata and FAIR metadata, which is one of the goal of the TRIPLE project. Also, to deal with this heterogeneity, we propose a data model for TRIPLE based on an ontology using schema.org vocabulary allowing to describe the different objects (Publications, datasets, projects and researchers profiles) in a standardized way taking into account the availability of data. Two versions of this ontology has been explored and proposed: a Full Ontology (FO), very precised and conceptual and a Light Ontology (LO), more flexible, realistic and less restrictive for the adopters (aggregators). Also, the TRIPLE platform will include Innovative Services, and we built this ontology by taking into account their needs. This ontology will be accessible in a OWL format online in the Huma-Num Gitlab.

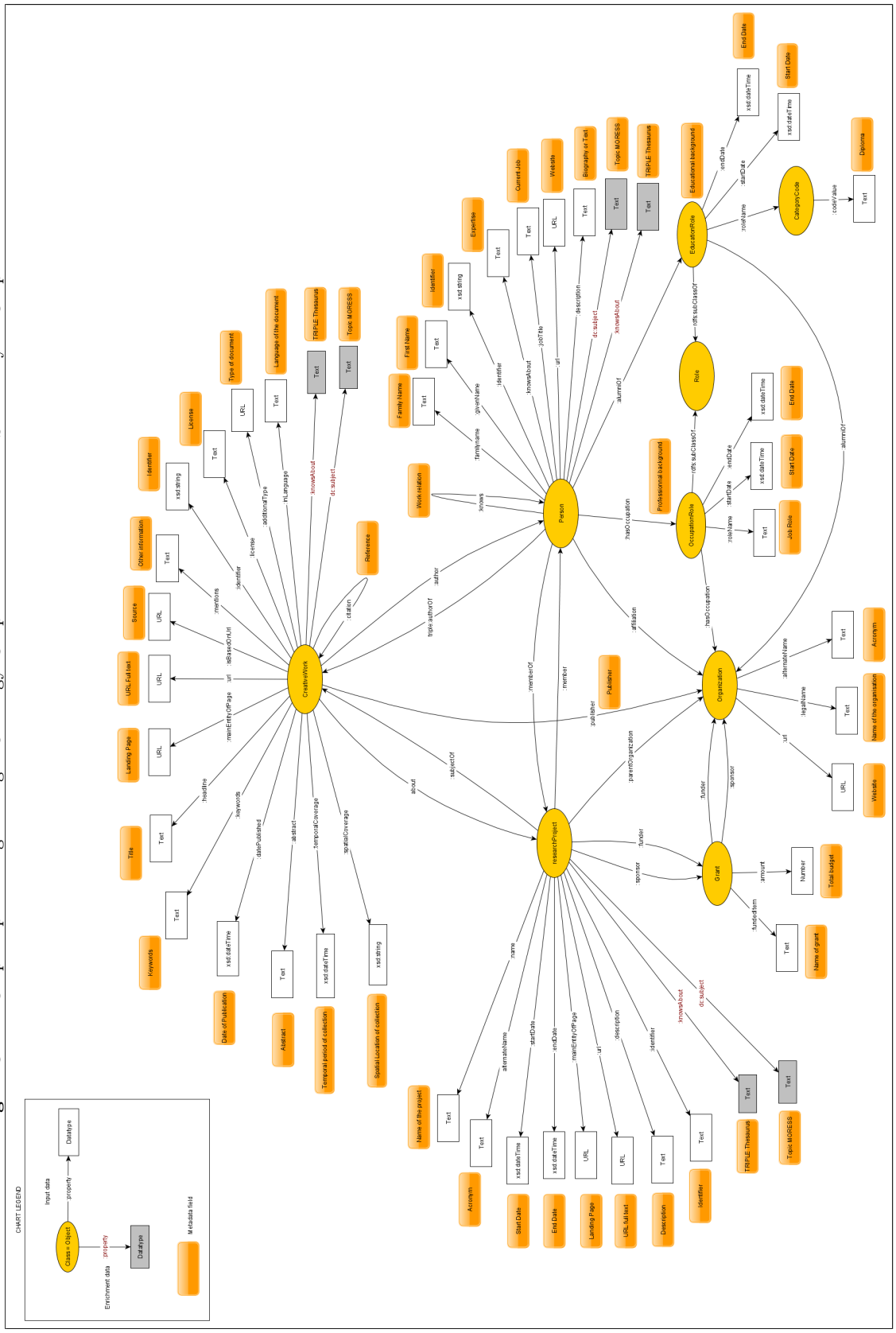



**Fig. 4.** Global Graph representing the Light Ontology. Graph realized under the Software yEd Graph Editor

# Table of Contents



## List of Figures



## References


1. Allemang, D., Hedler, J. : Semantic Web for the Working Ontologist : Effective Modeling in RDFS and OWL : Morgan-Kaufman, 2011
2. Aussenac-Gilles, N., Mothe, J. : Ontologies as Background Knowledge to Explore Document Collections, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), p. 129-142, 2004
3. Borst, P.: Construction of Engineering Ontologies for Knowledge Sharing and Reuse, Ph.D Dissertation, Tweente University,1997
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. Bunel, M., Capelli, L., Minel, J.L, Pouyllau, S. : Thinking about the Architecture of the Discovery Platform of the TRIPLE Project (2020). `https://halshs.archives-ouvertes.fr/halshs-02889863`
6. Pouyllau, S, Bunel, M., Capelli, L., Minel, J.L, . : "We" : a Proposal for the TRIPLE platform (2020). `https://halshs.archives-ouvertes.fr/halshs-02940860`
7. Caby, C. « Non obstante quod sunt monachi ». Être moine et étudiant au Moyen Âge. Quaderni di storia religiosa, 16 (Studia, studenti, religione), pp.45-81,(2009), halshs-00566289
8. Grubert, T.R. : A translation approach to portable ontology specifications, Knowledge Acquisition, 5 (2), p. 199-220, (1993)
9. Transforming Research through Innovative Practices for Linked interdisciplinary Exploration TRIPLE. The European discovery platform dedicated to SSH resources. Proposal number: 863420 (2018).
10. Dhombres, F., Jouannic, J., Jaulent, J., Charlet, J. : Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale, In S. DESPRÉS, M. CRAMPE, Eds., Actes des 21e Journées Ingénierie des Connaissances, Nîmes, France : Presse des Mines (2010)