A Model for Language Annotations on the Web

Frances Gillis-Webber[1][0000−0002−3740−5904], Sabine Tittel[2][0000−0003−4746−7604], and C. Maria Keet[3][0000−0002−8281−0853]

[1] Library and Information Studies Centre, University of Cape Town, South Africa; fran@fynbosch.com
[2] Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany; sabine.tittel@urz.uni-heidelberg.de
[3] Computer Science Department, University of Cape Town, South Africa; mkeet@cs.uct.ac.za

Abstract. Several annotation models have been proposed to enable a multilingual Semantic Web. Such models hone in on the word and its morphology and assume the language tag and URI comes from external resources. These resources, such as ISO 639 and Glottolog, have limited coverage of the world's languages and have a very limited thesaurus-like structure at best, which hampers language annotation, hence constraining research in Digital Humanities and other fields. To resolve this 'outsourced' task of the current models, we developed a model for representing information about languages, the **Mo**del for **L**anguage **A**nnotation (MoLA), such that basic language information can be recorded consistently and therewith queried and analyzed as well. This includes the various types of languages, families, and the relations among them. MoLA is formalized in OWL so that it can integrate with Linguistic Linked Data resources. Sufficient coverage of MoLA is demonstrated with the use case of French.

Keywords: Multilingual Semantic Web · Annotation · Language model

1 Introduction

Recent years have seen an appreciation of multilingualism in the global society, reflecting the increase of internet users, be it in spite of or thanks to the increase of English as *lingua franca* in global communication. This is on par with the trend toward the *both-and* attitude for cultural heritage, rather than an *either-or* of dominance and extinction. For example, it is European consensus that territorial varieties of languages need to be valorized and promoted, particularly online. International organizations emphasize the need for (culturally and) linguistically diverse local content to be published online; for a vitalization of multilingualism on the internet, see [30, 13–21]. Consequently, a fast-growing number of language resources have to be annotated, managed, and retrieved, not just for the few globally spoken languages, but also for the many local and regional languages.

For the Web, and the Semantic Web in particular, several proposals have been made to make it a *multilingual* Semantic Web, with solutions especially

for OWL ontologies and Linked Data in RDF; for a recent state of the art, see [8]. Language data expressed in RDF should be described in a principled way, and OntoLex-Lemon [8] by the W3C Ontology Lexicon Community Group is the *de facto* standard for representing the semantic, morphologic, and syntactic properties of lexical entries in linguistic resources. When modeling linguistic data using OntoLex-Lemon, the model requires the language to be defined using a URI. A lexical entry modeled[4] using OntoLex-Lemon is, e.g.:

```
:mola_mola  a    ontolex:LexicalEntry ;
   dct:language <http://id.loc.gov/vocabulary/iso639-2/eng> ,
                <http://lexvo.org/id/iso639-1/en> ;
   rdfs:label   "mola mola"@en ;
   ontolex:denotes dbr:Ocean_sunfish .
```

where the `dct:language` part is the focus of the paper. Recall that for RDF to be Linked Data, it should adhere to principles, among others: (1) "Use URIs as names for things", (2) "Use HTTP URIs so that people can look up those names", and (3) "When someone looks up a URI, provide useful information" (called a dereferenceable URI), using the RDF standard [2]. For the URI http://id.loc.gov/vocabulary/iso639-2/eng, although it is a persistent identifier for the language code *eng* from ISO 639 Part 2[5], information is not returned in RDF when navigating to this URI. Lexvo.org provides dereferenceable URIs for languages and mappings from Lexvo identifiers to only the ISO 639 language codes (Parts 1, 2, 3 and 5) [24]. Although ISO 639 is adequate for describing the world's main languages, when needing to assign a persistent identifier with a dereferenceable URI to a lesser-known language or dialect not included in ISO 639, an alternative catalogue has to be used. There are 7,865 language entries in ISO 639-3 [1], yet an estimated 3,000 to 10,000 languages are spoken in the world today, with some 150,000 extinct languages [10, 294-295]. Examples of alternative catalogues include Glottolog, Ethnologue, and MultiTree[6]. Glottolog is a comprehensive catalogue that provides reference information for language families, lesser-known languages, and dialects. Both Glottolog and MultiTree have persistent identifiers [17,18], but they do not have dereferenceable URIs. Ethnologue does not provide persistent identifiers to lesser-known languages and dialects, nor are there dereferenceable URIs.

Glottolog uses semantically underspecified 'broader/narrower than'-hierarchies, which is typical of Knowledge Organization Systems (KOSs). In order to account for languages and dialects, pseudo sub-groupings and names have been created that are, as the Glottolog developers also note, artificial. These shortcomings concern both under-resourced and well-resourced languages. For example, 'Loreto-Ucayali-Spanish' (Peru) is categorized under 'South Castilic', which is a sibling of 'Spanish' which, in turn, has a sub-language 'American Spanish'[7]: not only is 'South Castilic' not a sibling class of 'Spanish', its lan-

[4] For the sake of brevity, namespaces are assumed defined the usual way.
[5] ISO 639 is the International Standard for language codes [1].
[6] glottolog.org, www.ethnologue.com, multitree.org [05-03-2019].
[7] https://glottolog.org/resource/languoid/id/sout3200 [22-02-2019].

guage grouping is also questionable. Furthermore, it is not identifiable as being a pseudo-classification, hence it likely would—erroneously—be perceived as a legitimate classification by a non-expert. Another example is the subfamily 'Zulu-Xhosa' under Nguni, although no such classification exists. In fact, the Nguni group contains four languages on par with each other: isiZulu, isiXhosa, isiNdebele, and siSwati (all spoken in Southern Africa). Other language hierarchies and resources do not fare better; e.g., the alternate names given for isiXhosa in MultiTree[8] are archaic and hugely problematic, yet they have been propagated into Linked Data elsewhere, such as the US Library of Congress[9].

A KOS does not—and cannot—capture the intricacies of how 'languoids' (language family, sub-family, language, lect, or variant [17]) relate meaningfully: e.g., isiXhosa and isiZulu may be *sibling* languages where a language is *member of* a sub-family, Spanish and French *evolved from* Vulgar Latin and are *influenced by* Medieval Latin[10], and Afrikaans *evolved from* the Cape Dutch dialect that was a *dialect of* (old) Dutch. Not only is this an obstacle in the efforts to realize a multilingual Semantic Web but the underspecification of the subject domain and the lack of dereferenceable URIs for language tags also negatively impacts Humanities research pertaining to accurate language identification (cp. [29]). In addition, it hampers internationalization and localization.

In order to address these problems, we propose a model for representing information about languages: the '**Mo**del for **L**anguage **A**nnotation' (MoLA). MoLA provides a structured way for language annotation of objects on the Semantic Web. A modeler may include additional features of a languoid, such as its time period and geographic location for the period, as well as relate it to the language(s) it has evolved from, influences and has been influenced by. This enables more comprehensive RDF data about the languages of the world to be represented and, therewith, queried and analyzed. The model is formalized in OWL so as to achieve seamless integration with extant Linguistic Linked Data resources, and evaluated with competency questions and French as use case.

The remainder of the paper is structured as follows: Section 2 discusses related work, Section 3 describes MoLA, Section 4 revisits French and shows how MoLA is sufficiently expressive. We close with a discussion (Section 5) and conclusions (Section 6).

2 Related work

The most comprehensive resource for languages, particularly for under-resourced languages, is Glottolog, which we describe first. We discuss related works on KOSs and language models afterward.

[8] http://multitree.org/codes/xho [03-03-2019].
[9] http://id.loc.gov/authorities/subjects/sh85148822.rdf [03-03-2019].
[10] Language is constantly evolving. Influences by other languages due to cultural contact can result in lexical, phonetic and morphologic changes. The question when to characterize a language 'a' as 'being influenced' by a language 'b' depends on the granularity level of the analysis and is subject to discussion of linguists.

2.1 Glottolog as a KOS

Glottolog is a controlled vocabulary that provides a "comprehensive list of languoids" [17]. Each languoid is a concept as defined in SKOS [18, 195-196], where a SKOS concept "can be viewed as an idea or notion; a unit of thought" [25].

Each languoid as an (instance of a) concept is placed only once in the hierarchy (i.e., it does not have multiple inheritance) to represent genealogical relationships [11, 3]. The only SKOS relations Glottolog uses is `skos:broader` and `skos:narrower` [18, 195], i.e., there are no 'related term', 'use', or user-defined relations [25]. Glottolog also permits 'orphans', which are languoids that do not relate to another language [17], because too little information is known to reliably put it in the hierarchy.

Representing Glottolog's information. Based on the information provided by Glottolog about its system [17,18] and the data in the hierarchy, we have constructed a conceptual model of the information of its system. This is shown in Fig. 1 in Object-Role Modeling (ORM) notation [16], where the rounded rectangles are entity types, the smaller rectangles with a divider are the fact types (relationships), and dots and small lines on the relations are the constraints (mandatory and unique, respectively).

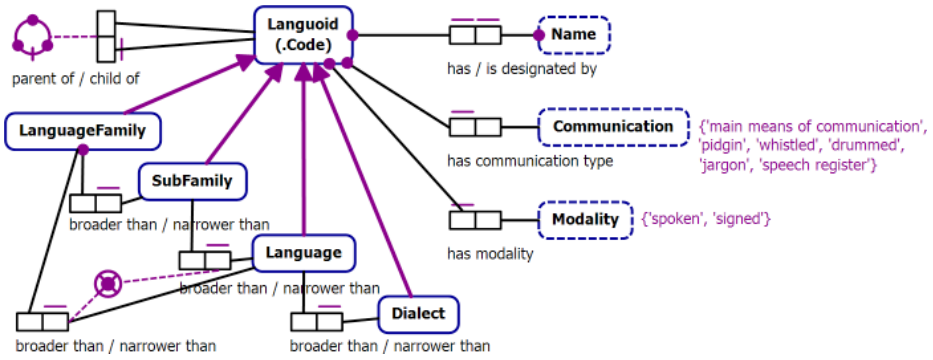

Fig. 1. Approximation of the conceptual model of Glottolog's system.

Use case: shortcomings of Glottolog's French. In its present state, Glottolog reveals major shortcomings with respect to the needs of linguists modeling data from the Romance languages, particularly regarding regional varieties and old language stages. For example, the categorization of varieties of French conflates diachronic and diatopic criteria within its hierarchies: Old French, the French spoken in the Middle Ages, is classified as a sibling of modern 'Central Oïl', Francoprovençalic, and Walloon (a French dialect).[11] Middle French however, the

[11] As sub-languoids of 'Oïl' (varieties that use an adaptation of the Vulgar Latin term *hoc ille* "this (is) it" as 'Yes'); note that Francoprovençalic is a non-Oïl language.

period following Old French, is classified four levels down into the branches and sub-branches of 'Central Oïl', together with 14 modern French varieties (including some of those spoken in the Americas) and, also, historical Anglo-Norman, spoken in England in medieval times. Other modern dialects are classified within other branches of 'Oïl'.[12] A selection of the 'broader/narrower than'-hierarchy of French varieties in Glottolog is shown in Fig. 2, which will serve as a means of comparison when we return to this case study in Section 4.

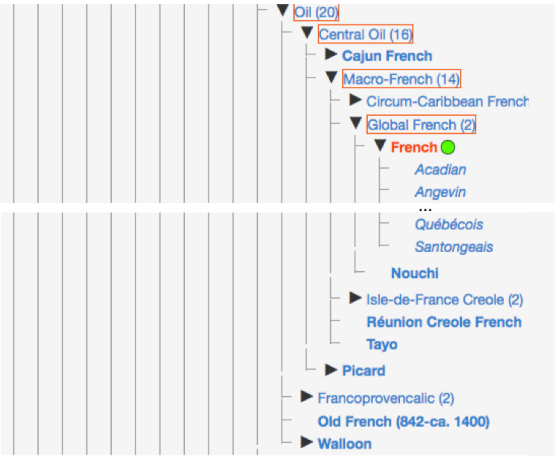

Fig. 2. Section of the 'broader/narrower than'-hierarchy of French (`stan1290`) in context in Glottolog.

2.2 Limitations of KOSs and other ontologies and models

Shortcomings with thesauri and KOSs are well-documented. For instance, the semantic underspecification of 'related term' RT (rather than a meaningful relation) means that one cannot query the system on, e.g., "what are the dialects of isiXhosa?" or "what are the languages between Vulgar Latin and modern-day French spoken in France?". Over the past twenty years, several proposals have been put forward to 'convert' a KOS to an ontology and add meaning in the process. Early works are, notably, the "rules as you go" proposal by Soergel et al. [28], who defines rules once a pattern is discovered during the manual stage of the conversion process. More recently, Kless et al. [22] proposed a method for converting thesauri and vocabularies more generally, but this is still a manual process and it was not evaluated beyond using illustrations from the popular

[12] Independent from diachronic issues, the hierarchy of modern French varieties also needs a revision (in line with "Most of the information on dialects in Glottolog [...] contains numerous errors and inconsistencies which we are aware of" [17]).

AGROVOC[13]. There are some attempts at automation, e.g. [5], but in the case of Glottolog, it would mean transferring the semantic errors into OWL, which does not help with querying and annotation needs.

There are several domain ontologies and models in the subject domain of languages. The OLiA ontologies [7] are domain ontologies for linguistic annotation at the word level and word fragment level of information, rather than the languages themselves that is needed for language tags and the management thereof. The NCS for linguistic task ontologies [6] are at the same level of detailed word level and morphology level of linguistic analysis. Hence, both are not applicable. The *Lemon* and OntoLex-Lemon models, as stated above, assume a suitable language tag is available, i.e., these models have 'outsourced' the language tags issue, and thus do not cover it themselves.

3 Designing MoLA

The development of MoLA followed a labour-intensive manual iterative bottom-up process with domain and knowledge engineering experts. The process adhered to the common main tasks of ontology development (as summarized and generalized in [27]) augmented with the explicit formulation of competency questions and the consideration of foundational ontology use. This is described in the next section, after which we present the content of MoLA.

3.1 Design approach

In order to demarcate the scope of the first version of the model that will improve sufficiently over the prevalent 'broader/narrower than'-hierarchies, we specify the following set of competency questions (CQs) for the model or (lightweight) ontology, as the case may be. The text in '[]' denotes a variable, meaning that it could take any subclass or individual classified into that class, as applicable.

1. Which languoids are dialects of [language]?
2. How many [languoids] does [language family] have?
3. Is a dialect a language?
4. Which types of languages have been classified?
5. Is a [languoid/language] divided into different time periods?
6. Does [language] have a region defined?
7. Which languages are spoken in [region]?
8. Is [language] the standard variety?
9. Is [language] in ISO 639?

When evaluating MoLA, these CQs must be answerable. We will revisit them further below, to test the efficacy of the proposed model.

We considered various principal approaches, methodologies, and methods for the development of the model (for a recent overview, see [19]):

1. reverse engineer the KOS (*in casu*, Glottolog) using a script;

[13] http://aims.fao.org/vest-registry/vocabularies/agrovoc [22-02-2019].

2. use a foundational ontology as the basis from which to start structuring the knowledge of the subject domain;
3. start from scratch with a 'clean slate', availing mainly of non-ontological resources for informal suggestions of names of classes, relations, and attributes.

We considered the first option unsuitable, because it would retain the underspecification and mis-classifications of the languages in the KOSs. For the second option, the generic, or at least top-domain, ontologies in the area of languages available are GOLD [12] and, to some extent regarding the design inspirations, DOLCE [23]. Due to the expected small size of the artifact, taking a top-down approach would unnecessarily clutter the ontology with classes and properties that would not be required, as only a very small fragment of a foundational ontology (FO) would be used. As such, we deemed more appropriate to indicate which elements of the model would bear a semantics as in one of those extant ontologies. A future, larger version of MoLA may include a module of a FO that will be aligned with equivalence and subsumption axioms. In addition, the competency questions are directed at ABox-level queries, rather than predominantly TBox-level, which suggests that the scope is more that of a knowledge base, ontology-based data access, and/or guidance for Linked Data. In that case, the artifact will not resemble an ontology in the principled sense, but rather a conceptual data model formalized in OWL for which the inclusion of a FO is atypical (see [15,19] for definitions and discussions thereof).

The third option amounts to creating the artifact from scratch. For this development process, we followed the process of scoping and then an iterative process cycling through the conceptualization, formalization, and evaluation stages. Besides consulting aforementioned resources, a domain expert also created content that has to be able to handle queries useful from that perspective (i.e., for digital humanities research). This domain input use case is depicted in Fig. 4. Subsequently, in a joint activity of the same domain expert (ST) and information and knowledge experts (FGW, CMK), we constructed a conceptual model using ORM notation which we formalized manually in OWL. Collaborative software used were mainly WebProtégé and GitHub, as well as the standalone tools Norma (for VS2017) and the Protégé v5.x desktop version.

3.2 Content

The core idea of Glottolog's languoid is reused in MoLA, although the conceptual model in Fig. 3 and the subsequent OWL file of MoLA is more comprehensive. Importantly, several relations between languages have now been included, effectively adding the semantics that is typically underrepresented in KOSs, as well as the addition of basic time and space properties.

There are different definitions of 'language' in the literature but the consensus can be described as follows: Language is a complex and heterogeneous but structured system of communication used within a community of speakers. Within this system, a number of varieties—also called lects—reflect diatopic aspects referring to geographic areas (regional varieties, dialects, patois), diaphasic aspects

referring to the communicative context (formal or informal style, technical language), and diastratic aspects referring to the social classes (sociolect, idiolect, youth language) [9]; cp. [4, 14]. These thus resulted in the most relevant classes in MoLA. The ABox axioms are primarily class membership assertions, e.g., `Language(vulgar_latin)`. `Lect` and its sub-classes are added, for it would be expected by a sociolinguist. Instead of broader/narrower, languages may now be member of a language family, or dialects may be member of a dialect cluster. RBox axioms were mostly domain and range axioms and inverses. To permit inclusion of languoids for which only partial information is known, few hard constraints have been enforced.

Some entities seem to operate at different levels of granularity, such as that a language may refer to a collection of dialects at the finer-grained level of analysis. This distinction is reflected in MoLA with the collections.

Because the model needs to be used in praxis and record data about individual languoids, the other salient feature of the model is that there are data properties and data types, such as the link to ISO codes, for compatibility with other language resources.

Notable distinctions with Glottolog (recall Fig. 1) and other sources are:

- Instead of the broader/narrower relation, there is the proper subsumption relation and separate meaningful relations, such as *influenced by*;
- a language can be in more than one language family;
- the uniqueness of both languoid code and language name is no longer required, therewith more easily permitting one languoid to have multiple names and labels;
- a language family or a lect can be associated with ≥ 0 regions and periods;
- relations to/with other languages can also be represented explicitly and relatively meaningfully, including the influence on another language, and the evolution of a language;
- a languoid can be associated with language codes from ISO 639;
- a language can be associated with one or more custom language tags, as defined by IETF's BCP47, accounting for both varying regions and periods.

The translation from the ORM diagram to OWL faced only one real obstacle: time periods ought to be represented with data type `gYear`, but this XML datatype is not supported by the OWL standard. Therefore, it was encoded as an xml:string. Acyclicity on `evolvedFrom` and `influences` also cannot be represented in OWL.

Considering the conceptual model and the CQs, the result is a model formalized in OWL, with the characteristics of what may be called an "application ontology", which is available at https://ontology.londisizwe.org/mola.

3.3 Validation

As first pass of validation, we describe an example of MoLA's use and the CQs; the use case with French is deferred to the next section.

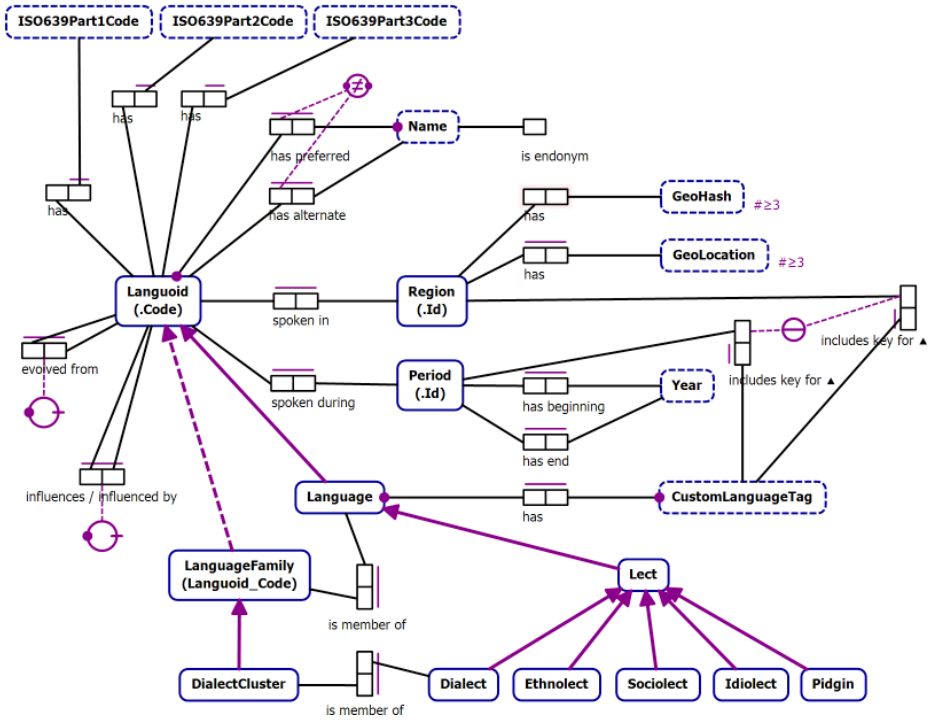

Fig. 3. The Model for Language Annotation, MoLA.

Illustration of MoLA usage. To demonstrate that MoLA works for a modern language stage, we use Spanish as spoken in Cuba (refer to [3, 23–25; 102–106] for an overview of South American Spanish): it is 'Caribbean Spanish' spoken in Cuba since the first half of the 16th century, it is influenced by Spanish of the Canary Islands, French, the indigenous language Taíno, and West African languages of the slaves (Niger-Congo languages?), it evolved from Spanish and influences South-American Spanish, and it is the official language. Using MoLA, a fragment of the encoding is as follows:

```
:cuban_spanish
    a                   mola:Dialect ;
    rdfs:label          "Cuban Spanish"@en,
                        "Español Cubano"@es-CU ;
    dct:language        :cuban_spanish;
    mola:inFamily       :spanish ;
    mola:isMemberOf     :caribbean_spanish ;
    mola:inPeriod       :cuban_period ;
    mola:inRegion       :cuba_region ;
    mola:evolvedFrom    :spanish ;
    mola:influencedBy   :canary_island_spanish , :spanish,
                        :french , :taino ,
```

```
13                              :westafrican_languages ;
14      mola:influences         :spanish , :southamerican_spanish .
```

Representing the current knowledge of the languoids with MoLA may suffice for some users, but it would be only a prerequisite for Digital Humanities end-users. For instance, when text documents are annotated with MoLA, one could retrieve all `caribbean_spanish` text documents for some NLP task, search the web for websites in isiXhosa dialects, or search for Medieval English texts without having to specify the exact start or end year. As such, it could assist semantic search by providing additional parameters in the query, hence better narrowing down the information request.

Competency questions revisited. Revisiting the CQs, all questions are answerable, with the exception of CQ 8 (which is planned for a next version, see also the discussion in Section 5). CQ 3, 5 and 7 are shown here, with the remaining CQs answered in the online supplementary material at [13]. For **CQ 3**, this can be answered with the following query (in SPARQL-OWL shorthand notation) $\alpha \leftarrow$ SubClassOf(Dialect Language) where α is the answer, being 'yes'. For **CQ 7**, this can be answered with the query $\alpha \leftarrow$ Type(Languoid ObjectSomeValuesFrom(inRegion cuba_region)), which will retrieve, at least, `cuban_spanish`. For **CQ 5**, this can be answered with the SPARQL query ASK { :cuban_spanish mola:inPeriod ?any } where the result will return 'True'.

Note that the ontology-as-knowledgebase is not fully populated with language data, so an answer may be empty. Most CQs are knowledge base queries, in fact, not TBox queries. While this may seem disappointing from an ontology development viewpoint, it merely highlights the prospective aims where the language annotations are needed.

4 Use case: Structuring French languoids

As briefly mentioned in Section 2, there are various shortcomings of how French has been represented with respect to the state of affairs scientifically. Therefore, we deem a remodeling of French with its historical language stages and dialects necessary to meet the needs of linguists. To put the remodeling on solid historico-linguistic ground, we consider the formation of the French language and the development of its spoken and written varieties, as visualized in Fig. 4 and briefly explained in the following. We identified five language periods of French.

period I Old French: 842 AD (*Serments de Strasbourg*: 'formation deed' of France [4, 183–189]) – ca. 1350 (major grammatical changes),
period II Middle French: ca. 1350 – ca. 1500,
period III French of the Renaissance: ca. 1500 – 1605 (influences of the Reformation, the Humanists, and the Renaissance [21, 89]),
period IV Classical and neo-classical French: 1605 (F. de Malherbe was called to the court of Henri IV [21, 116]) – 1789 (French Revolution),
period V 1789 – today.

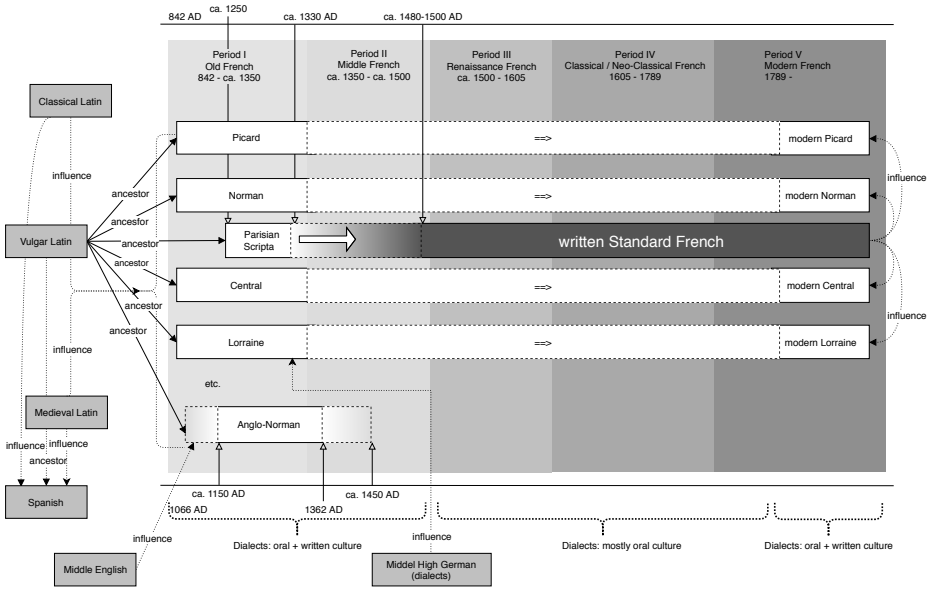

Fig. 4. Diachronic diagram of French.

Primary dialects and the emergence of French as a standard. During the Old French period, the diagram shows the emergence of the primary dialects, Old French scriptae respectively[14], that were the result of the Romanization process and derived from Vulgar Latin, the (almost exclusively) spoken form of Latin. The dialects and the standard French have been influenced by written Classical Latin and also Medieval Latin from the beginning of the Romanization until today [26, 91-93; 118f.; 142]. Examples of Old French dialects are ancient Picard, ancient Norman, ancient Lorraine or Anglo-Norman. Anglo-Norman was used in England from mid 12$^{\text{th}}$ century until mid 15$^{\text{th}}$ century [20, 5–19; 57f.; 92].

Next to the Old French scriptae, a Parisian scripta started to emerge around 1250 and around 80 years later it started to spread as a standardized written variety of Old-/Middle French, gradually replacing the regional scriptae; this process was completed around 1480–1500 [4, 203–211]. The third period then witnesses the constitution of French as a national language [21, 89].

The relations visualized in Fig. 4 and explained above can be described as follows for ancient **Lorraine**: it is a dialect included in the notion of 'Old French', spoken in time period I in the region of Lorraine (north-eastern France), it evolved from Vulgar Latin, and is related to other Old French scriptae (e.g., Picard, Norman, members of the same language family), it is influenced by Clas-

[14] The term for the written representation of the spoken dialects of Old French [4, 206].

sical Latin, Medieval Latin, and (dialects of) Middle High German.[15] In Turtle notation with MoLA, one obtains:

```
:old_french_lorraine
    a                   mola:Dialect ;
    rdfs:label          "Lorraine"@en ;
    dct:language        :old_french_lorraine ;
    mola:isMemberOf     :old_french ;
    mola:inPeriod       :french_period_one ;
    mola:inRegion       :old_french_lorraine_region ;
    mola:evolvedFrom    :vulgar_latin ;
    mola:influencedBy   :classical_latin, :medieval_latin,
                        :middle_high_german .

:french_period_one
    a                   mola:Period ;
    rdfs:label          "Old French Period"@en ;
    mola:hasBeginning   "842"^^xsd:string ;
    mola:hasEnd         "1350"^^xsd:string ;
    mola:duration       "508"^^xsd:int .

:old_french_lorraine_region
    a                   mola:Region ;
    mola:hasCoordinate  :old_french_lorraine_region_coord1 ,
                        :old_french_lorraine_region_coord2 ,
                        :old_french_lorraine_region_coord3 ,
                        :old_french_lorraine_region_coord4 ;
    rdf:_1              :old_french_lorraine_region_coord1 ;
    rdf:_2              :old_french_lorraine_region_coord2 ;
    rdf:_3              :old_french_lorraine_region_coord3 ;
    rdf:_4              :old_french_lorraine_region_coord4 .

:old_french_lorraine_region_coord1
    a                   mola:GeographicCoordinate ;
    geo:lat             "4.91473"^^xsd:decimal ;
    geo:long            "49.62686"^^xsd:decimal .

# Due to space constraints, other coordinates are not shown.

:old_french         a   mola:LanguageFamily .
:vulgar_latin       a   mola:Language .
:classical_latin    a   mola:Language .
:medieval_latin     a   mola:Language .
:middle_high_german a   mola:Language .
```

[15] I.e., Moselle and Rhine Franconian for which a thorough revision on Glottolog is advised as well, see https://glottolog.org/resource/languoid/id/fran1268 [24-02-2019].

5 Discussion

Designing MoLA exhibited several main challenges that had to be resolved, one of which was principally a knowledge engineering issue, the other that of languages. Regarding the former, this first issue concerned `LanguageFamily` and `Language`, which surfaced for Old French, and induced related questions on dialect clusters and language: it is not the case that something is ontologically two different kinds of things *at the same time*, but it depends on (1) what a term denotes, in particular whether it is monosemous or polysemous, and (2) the level of granularity of analysis of the entity. The modeling issue exhibited both, which was due to mixing levels of granularity. For instance, the term 'Old French' has several meanings: (i) it is used to refer to the *language* spoken by the people in the northern part of what is now France, (ii) it is the umbrella term for a collection of dialects, (iii) it is used to precisely designate the intersection of (lexemes, phonemes, syntactic structures of) the distinct varieties that are part of the collection. Old French is a language and a member of the family of the Romance languages. At the same time, however, Old French is a language family consisting of the distinct dialects such as Picard and Norman. And then, also Picard can be seen as a language system with a number of varieties such as the ones spoken in Artois and Santerre which makes it a language family. Regarding the latter, the question "how to define 'language' and 'language family'?" is inescapable, and the consensus approach was taken. For the notion of the 'ancestor' relation between languages, there was no consensus, however the pseudo-synonymity of the verbs *to evolve* and *to derive* used in the literature led to the decision to only introduce one property, i.e., `evolvedFrom` to MoLA.

MoLA enables a modeler or annotator to define both periods and regions for a languoid, reflecting its diachronicity. A custom language tag, encoded using a pattern [14], can then be associated with languoid and period or region or both, which is more comprehensive than the standard ISO 639 language codes. Not only does MoLA provide dereferenceable URIs with persistent identifiers, it can also be queried, returning useful information about that languoid. This thus means that it is, by design, amenable to accommodate new languages and the identification and recording of extinct languages. In this respect, MoLA is apolitical. It is envisioned for a version 2 to accommodate also what Glottolog calls "orphans" (i.e., linguists do not know where to classify it yet), and the societal or political status of a language, with notions such as official, standard, minority, and dominant languages, a country's *lingua franca*, and contested languoids.

Note that MoLA does facilitate diachronic naming of languoids by availing of the `period` property, for when there are alternate names of languages over different periods. For instance, the official Dutch used to be called *Algemeen Beschaafd Nederlands* 'General Civilised Dutch' until the 1970s, which is now called *Standaardnederlands* 'Standard Dutch'. This may also resolve the aforementioned problem with isiXhosa in MultiTree (see fn. 8), as one of the older names of isiXhosa listed there would have one disciplined, at the very least, if used in present-day South Africa. Alternate *current* names can be specified as

well, but, in version 1, only as preferred and alternate. Model extensions in this direction are possible and planned.

6 Conclusions

The paper presented a proposal for relatively comprehensive and semantically meaningful language annotation tags, well beyond the extant lists and structured resources of basic Knowledge Organisation Systems. This Model for Language Annotation, MoLA, whilst backward-compatible with these systems, allows a user to specify more languages, lects, and language and dialect families ('languoids'), as well as some of their properties, such as the region and time period they are or were spoken in, and relations among the languoids, such as which language is evolved from or is influenced by which other language. In conjunction with the *de facto* standard word-level annotation models, or on its own, it enables more detailed querying of language information and MoLA-annotated documents and objects on the Web that can be useful for Digital Humanities.

MoLA has been demonstrated to sufficiently represent the complexity of Old French. Future work includes populating it with more languoids and their properties as well, extending the model with further possible information about languages. Given the Linked Data direction of applicability of language tags, MoLA is intended to be published in the Linguistic Linked Open Data cloud.

References

1. Language codes - ISO 639 (nd), https://www.iso.org/iso-639-language-codes.html
2. Berners-Lee, T.: Linked Data (2009), https://www.w3.org/DesignIssues/LinkedData.html
3. Berschin, H., Fernández-Sevilla, J., Felixberger, J.: Die spanische Sprache. Georg Olms Verlag, Hildesheim / Zürich / New York (2012)
4. Berschin, H., Goebl, H.: Französische Sprachgeschichte. Georg Olms Verlag, Hildesheim / Zürich / New York (2008)
5. Cardillo, E., Folino, A., Trunfio, R., Guarasci, R.: Towards the reuse of standardized thesauri into ontologies. In: Proc. of WOP'14. CEUR-WS, vol. 1302, pp. 26–37 (2014)
6. Chavula, C., Keet, C.M.: An orchestration framework for linguistic task ontologies. In: Proc. of MTSR'15. CCIS, vol. 544, pp. 3–14. Springer (2015), 9-11 Sept., 2015, Manchester, UK
7. Chiarcos, C., Sukhareva, M.: OLiA – ontologies of linguistic annotation. Semantic Web Journal **6**(4), 379–386 (2015)
8. Cimiano, P., McCrae, J.P., Buitelaar, P.: Lexicon model for ontologies: Community report. Final community group report, 10 may 2016, W3C (2016), https://www.w3.org/2016/05/ontolex/
9. Coseriu, E.: 'Historische Sprache' und 'Dialekt'. In: Göschel, J. (ed.) Dialekt und Dialektologie. Ergebnisse des Internationalen Symposions "Zur Theorie des Dialekts". Marburg/Lahn, 5.–10. Sept. 1977. pp. 106–122. Franz Steiner Verlag (1980)
10. Crystal, D.: The Cambridge Encyclopedia of Language. Cambridge University Press, Cambridge (2010)

11. Dimitrova, V., Fäth, C., Chiarcos, C., Renner-Westermann, H., Abromeit, F.: Interoperability of Language-related Information: Mapping the BLL Thesaurus to Lexvo and Glottolog. In: Proc. of LREC 2018. pp. 4555–4561. ELRA, Miyazaki, Japan (May 7-12 2018)
12. Farrar, S., Langendoen, D.T.: A linguistic ontology for the semantic web. In: GLOT International. 3, vol. 7, pp. 97–100 (2003)
13. Gillis-Webber, F., Keet, C.M., Tittel, S.: A model for language annotations on the web: Supplementary material (2019), https://ontology.londisizwe.org/mola/article/2019-kgswc-supplementary-material
14. Gillis-Webber, F., Tittel, S.: The Shortcomings of Language Tags for Linked Data when Modelling Lesser-Known Languages. In: Proceedings of LDK2019. Leipzig (May 20-23 2019)
15. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, chap. 1, pp. 1–17. Springer (2009)
16. Halpin, T., Morgan, T.: Information modeling and relational databases. Morgan Kaufmann, 2nd edn. (2008)
17. Hammarström, H., Haspelmath, M., Forkel, R.: Glottolog 3.3. About languoids (2018), https://glottolog.org/glottolog/glottologinformation, accessed: 2019, February 17
18. Hellmann, S., Stadler, C., Lehmann, J.: Linked data for linguistic diversity research: Glottolog/Langdoc and ASJP online. In: Linked Data in Linguistics, pp. 191–200. Springer (2012)
19. Keet, C.M.: An introduction to ontology engineering, Computing, vol. 20. College Publications, UK (2018), 334p
20. Kibbee, D.: For to Speke Frenche Trewely. John Benjamins Publishing Company, Amsterdam / Philadelphia (1991)
21. Klare, J.: Französische Sprachgeschichte. Klett, Stuttgart (1998)
22. Kless, D., Jansen, L., Lindenthal, J., Wiebensohn, J.: A method for re-engineering a thesaurus into an ontology. In: Proc. of FOIS'12. pp. 133–146. IOS Press (2012)
23. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Ontology library. WonderWeb Deliverable D18 (ver. 1.0, 31-12-2003). (2003), http://wonderweb.semanticweb.org
24. de Melo, G.: Lexvo.org: Language-related information for the Linguistic Linked Data cloud. Semantic Web **6**(4), 393–400 (August 2015)
25. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System reference: W3C recommendation 18 August 2009 (2009), https://www.w3.org/TR/2009/REC-skos-reference-20090818/, accessed: 2019, February 17
26. Rickard, P.: A history of the French language. Hutchinson University Library, London (1974)
27. Simperl, E., Mochol, M., Bürger, T.: Achieving maturity: the state of practice in ontology engineering in 2009. International Journal of Computer Science and Applications **7**(1), 45–65 (2010)
28. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering thesauri for new applications: the AGROVOC example. Journal of Digital Information **4**(4) (2004), http://journals.tdl.org/jodi/article/view/jodi-126/111
29. Tittel, S., Chiarcos, C.: Historical Lexicography of Old French and Linked Open Data: Transforming the Resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon. In: Proc. of LREC 2018. GLOBALEX Workshop 2018, Miyazaki, Japan, 2018. pp. 58–66. Paris (ELRA) (2018)
30. Vannini, L., Le Crosnier, H.: Net.lang. Towards the multilingual cyberspace. C & F Éditions (2012)