The Shortcomings of Language Tags for Linked Data When Modeling Lesser-Known Languages

Frances Gillis-Webber
Library and Information Studies Centre, University of Cape Town, South Africa
http://www.dkis.uct.ac.za
fran@fynbosch.com

Sabine Tittel
Heidelberg Academy of Sciences and Humanities, Germany
http://www.deaf-page.de
sabine.tittel@urz.uni-heidelberg.de

Abstract

In recent years, the modeling of data from linguistic resources with Resource Description Framework (RDF), following the Linked Data paradigm and using the OntoLex-Lemon vocabulary, has become a prevalent method to create datasets for a multilingual web of data. An important aspect of data modeling is the use of language tags to mark lexicons, lexemes, word senses, etc. of a linguistic dataset. However, attempts to model data from lesser-known languages show significant shortcomings with the authoritative list of language codes by ISO 639: for many lesser-known languages spoken by minorities and also for historical stages of languages, language codes, the basis of language tags, are simply not available. This paper discusses these shortcomings based on the examples of three such languages, i.e., two varieties of click languages of Southern Africa together with Old French, and suggests solutions for the issues identified.

2012 ACM Subject Classification Computing methodologies → Language resources; Information systems → Dictionaries; Information systems → Semantic web description languages; Information systems → Graph-based database models; Information systems → Resource Description Framework (RDF); Software and its engineering → Interoperability; Information systems → Multilingual and cross-lingual retrieval; Computing methodologies → Information extraction; Computing methodologies → Artificial intelligence

Keywords and phrases language codes, language tags, Resource Description Framework, Linked Data, Linguistic Linked Data, Khoisan languages, click languages, N|uu, ||'Au, Old French

Digital Object Identifier 10.4230/OASIcs.LDK.2019.4

Acknowledgements We would like to thank the reviewers for helpful comments and insightful feedback.

1 Introduction

The publication of language data on the Web as Resource Description Framework (RDF), and according to Tim Berners-Lee's Linked Data principles[1], has contributed to the emergence of a multilingual web of data. Publishing language resources as Linked Data allows for language resources to be exploited with the benefits of structural interoperability (same format and query language leading to cross-resource access), conceptual interoperability (shared standard vocabularies), accessibility (via standard Web protocols), and resource integration (via linked resources) [6].

[1] https://www.w3.org/DesignIssues/LinkedData.html [10-01-2019].

© Frances Gillis-Webber and Sabine Tittel;
licensed under Creative Commons License CC-BY
2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 4; pp. 4:1–4:15
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

After a brief introduction to RDF and Linked Data, particularly in the context of linguistic resources, as well as language codes and language tags (Section 1), we present the challenge addressed in this paper: finding solutions for the shortcomings of language tags when identifying near-extinct and historical languages (Section 2), and we do so by modeling data from three languages, e.g., two click varieties from the language family previously referred to as 'Khoisan', and Old French (Section 3). The paper concludes with a discussion of the findings (Section 4) and directions for future work (Section 5).

1.1 RDF and (Linguistic) Linked Data

RDF is the standard data model for resources of the Semantic Web [9]. It expresses data as *subject-predicate-object* triples to facilitate data interchange on the web. Each *subject* and *object* is a node; the *predicate* forms a relation (edge) between two nodes. The *subject* can be a URI (Uniform Resource Identifier) or a blank node, the *predicate* can only be a URI, and the *object* can be a URI, blank node or a literal (described as a string), see [9, 3].

Linked Data (LD) can be defined as the «set of best practices for publishing and connecting structured data on the Web», and it builds on the RDF data model using HTTP (Hypertext Transfer Protocol) URIs [35, 4-12]. The LD principles have been adapted in many fields, including linguistics, where it has led to the creation of numerous datasets published as Linguistic Linked Open Data (LLOD)[2]: lexicons, annotated corpora, dictionaries, etcetera ([4, 24]). The model that has become the *de facto* standard for describing linguistic resources is the OntoLex-Lemon vocabulary[3] [26, 587]. The focus within this field lies on well-resourced languages and, in particular, on their modern stages, with a small number of examples of linguistic resources documenting low-resourced languages (e.g., [13, 27, 15]) and also historical language stages (e.g., [32, 7, 22, 31, 2]).

1.2 Language codes → language tags

To use unique codes for the identification of languages is necessary for any environment that follows BCP 47 [28]. Examples include language identification in RDF and XML documents (the latter using the `xml:lang` attribute), and institutions such as language repositories, e.g., the Open Language Archives Community (OLAC) and the World Atlas of Language Structures (WALS).[4] A unique language code is able to disambiguate the case when one language name refers to several languages, and one language has several names.

A language code «represents one or more language names, all of which designate the same specific language» [19]. The International Organization for Standardization (ISO) provides a standard for language codes: ISO 639 with Parts 1–3. In principle, the language codes in each part «are open lists that can be extended and refined», and a Registration Authority nominated by ISO maintains each part [12]. ISO 639-1 provides a two-letter code and it is a subset of ISO 639-2, which provides a three-letter code allowing for more languages to be represented. Both ISO 639-1 and ISO 639-2 represent major languages that are most frequently expressed in the world's literature ([12, 18]). The individual languages in ISO 639-2 are in turn a subset of those in ISO 639-3 that aims «to give as complete a listing of languages as possible» [12]. The types of languages covered include living, extinct, ancient, historic and constructed languages; their scope can either be an individual language or a macrolanguage, and the modality is spoken, written or signed ([12, 18, 19, 20]).

[2] `http://linguistic-lod.org/` [26-12-2019].

[3] `https://www.w3.org/2016/05/ontolex/` [31-12-2018].

[4] `http://www.language-archives.org/`; `https://wals.info/` [15-03-2019].

For individual languages, only varieties which are considered to be distinct languages are represented in ISO 639-3, with any dialects encompassed within the language code of that language. The language code «represents the complete range of all the spoken or written varieties of that language, including any standardized form» [19]. A macrolanguage code represents a cluster of language varieties. Macrolanguages differ from language collections in that for the former, the languages must be deemed very closely related, and for the latter, there can be a loose relation, but there should be some connecting feature, be it historical, geographical, or a linguistic association [28, 33]; language collections are only represented in ISO 639-2, and macrolanguages are only represented in ISO 639-3 [19].

A language tag is similar in concept to a language code, except the latter can be used in any discipline, and the former is intended for the internet community. The scope of a language tag is defined by IETF's BCP 47. BCP 47 is a document which specifies Best Current Practice for tags for identifying languages, and the language in question is able to be refined further from the ISO 639 language code ([12]; [28, 1-4]; [21]). Language tags are of the form: *language-extlang-script-region-variant-extension-privateuse*, comprised of one or more sub-tags, each separated by a hyphen; *language* is the shortest language code from ISO 639, and the remaining sub-tags are distinguished from each other «by length, position in the tag, and content» ([21]; [28, 4]).

1.3 Language codes for linguistic resources

The OntoLex-Lemon specification requires each linguistic resource, be it a lexicon, a lexical entry, or a lexical concept, to be identified using a URI to the relevant ISO 639 code, with RDF requiring each string literal in an *object* to be 'language-tagged'.[5]

A language code (or tag) is thus used in the following scenarios:

1. to identify a lexicon:
   - when a triple with the predicate `dct:language`[6] is declared: this is to the URI of an ISO 639 language code [8];
2. to identify a lexical entry:
   - same as (1);
3. for the language tagging of string literals:
   - this is a language tag, which, in the absence of additional sub-tags, is an ISO 639 language code [9].

A lexical entry in RDF, described using OntoLex-Lemon and serialized in Turtle[7], can be modeled as follows:

```
@PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@PREFIX dct:     <http://purl.org/dc/terms/> .
@PREFIX rdfs:    <http://www.w3.org/2001/02/rdf-schema#> .

:entry/en-n-bile a ontolex:LexicalEntry , ontolex:Word ;
   lexinfo:partOfSpeech   lexinfo:Noun ;
   dct:language           <http://id.loc.gov/vocabulary/iso639-2/en> ,
                          <http://lexvo.org/id/iso639-1/en> ;
   rdfs:label             "bile"@en ;
```

[5] `https://www.w3.org/2016/05/ontolex/#conventions-in-this-document` [10-01-2019].

[6] Beyond RDF, OntoLex-Lemon, and DublinCore (dct) vocabulary, we use classes and properties of LexInfo, RDFS, SKOS, and DBpedia, see the respective URLs within the code examples.

[7] Terse RDF Triple Language, an easy to read serialization of RDF statements, `http://www.w3.org/TR/turtle/` [11-01-2019].

```
11     ontolex:canonicalForm  :entry/en-n-bile#lemma ;
12     ontolex:sense          :entry/en-n-bile#sense1 ;
13     ontolex:evokes         :concept/000000001 .
```

Where:

- Point 2 is demonstrated in Line 8-9: the applicable language codes for the lexical entry, from ISO 639-2 and ISO 639-1 respectively, are indicated as ‘English’.
- Point 3 is demonstrated in Line 10: the language of the literal “bile” is specified with the ISO 639-1 code for English.

2 The shortcomings of language tags

The ISO 639 standard list includes more than 6,900 language codes[8] but it neither covers all the world’s languages nor all historical language stages of the languages. This is problematic when modeling under-resourced or extinct languages for which a language code does not exist. To the best of our knowledge, this problem has not been properly addressed in the literature. A recent email thread in the W3C Semantic Web forum[9] expressed the opinion to do away with language tags altogether, but there was not shared consensus on this point.

Chiarcos and Sukhareva [7] show the conversion of legacy data from dictionaries of the historical language stages of Germanic languages (Old Saxon, Old High German, Old Norse, etc.) and find the following compensation for the lack of language codes within ISO 639: they preserve the original language abbreviations of the dictionary resource and extract «all language identifiers, and by a hand-crafted mapping from the original abbreviations», ISO 639-3 codes are assigned where possible [7, 44b]. The language URIs are represented using *lexvo* [10], but «[u]nfortunately, many abbreviations could not be resolved against lexvo, in particular, this included hypothetical forms for reconstructed historical language stages, e.g., Proto-Germanic.» They conclude that the extension of existing terminologies with respect to historical language stages is a great desideratum [7, 44b]. Their approach results in code such as `lemon:language "ae."@deu`, with ‘ae.’ being the German abbreviation for *Altenglisch* in the dictionary resource [23], see [7, 44b], and ‘deu’ being the ISO 639-3 language code for Standard German.[10]

The same approach has been taken by Declerck et al. for the transformation of the data from the *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ)[11] into LD [11]: in the code sample given at [11, 347], the language tag for the Bavarian language is modeled as a literal: `bar"^^xsd:string"`, which raises the question why it is not given in the form of `@bar`, ‘bar’ being the ISO 639 code for Bavarian.[12] One might speculate that this could serve as a means to distinguish the language documented in the WBÖ (Bavarian varieties spoken in Austria) from Bavarian spoken in Bavaria; however, the problem is not addressed in the paper.

Amongst the findings of Tittel and Chiarcos [32] is the fact that due to the lack of appropriate language codes, the problem of modeling the different dialectal forms of lexemes in linguistic resources of Old French is still unsolved: for the conversion of the data of

[8] According to the table in `https://de.wikipedia.org/wiki/ISO_639` [10-01-2018].

[9] Language-tagged strings Re: Towards easier RDF: a proposal [Electronic mailing list, 23-26 November] 2018, `https://lists.w3.org/Archives/Public/semantic-web/2018Nov/thread.html#msg90` [01-01-2019].

[10] `https://iso639-3.sil.org/code/deu` [11-01-2019] (639-1: ‘de’; 639-2/B: ‘ger’).

[11] `https://wboe.oeaw.ac.at/` [11-01-2019].

[12] `https://iso639-3.sil.org/code/bar` [11-01-2019].

the *Dictionnaire étymologique de l'ancien français* (DEAF, [1]) following the Linked Data paradigm, the researchers established that all graphical variants of a given Old French lexeme could only be identified by ISO 639-3 code 'fro'[13] for overall Old French. This meant that information originally included in the linguistic resource such as 'Anglo-Norman' or 'medieval Lorraine' scripta[14] – information that is very valuable for the research of Old French dialects – would be excluded from the language description when converted to Linguistic Linked Data. To solve this problem, [32, 65] propose to define the code 'fro' in ISO 639-3 as a macrolanguage and to register the Old French dialects as varieties associated to 'fro'. (There had been an attempt to include varieties of historic languages within ISO 639-6, but this Part was withdrawn in 2014.[15])

Bellandi et al. [2] discuss the modeling of linguistic data from Old Occitan (a Romance language spoken during the Middle Ages in what is today southern France) and other languages using OntoLex-Lemon. To code their Old Occitan lexemes, they use the tag 'aoc': `lemon:writtenRep "canabo"@aoc` [2, 4]. One rightly assumes that this is the ISO 639-3 code 'aoc', however, 'aoc' represents the Pemon language of the Cariban language family, a language in Venezuela.[16] The correct ISO 639 code for the language is 'pro' (= Old Provençal, the former term for the language)[17], and presumably 'aoc' simply is an abbreviation for French *ancien occitan*. Their handling of the use of codes is illegal: the definition of a language tag using the '@' sign and a language code must be BCP 47 compliant to be valid.[18] [2] do not address this issue, nor do they address the issue of creating their own language codes.

We conclude that new language codes need to be created, in a way that adheres to current standards and best practices of language identification. The objective of this paper is to contribute to the discussion of this problem. On the basis of three example languages, we will propose solutions to meet the requirements of the languages discussed.

The following languages serve as our examples:

1. N|uu and ||'Au: two dialects from N||ng, a critically endangered non-Bantu click language in Southern Africa, that are both near-extinct [30, 7].
2. Old French: the ancestor of modern French, spoken during the Middle Ages.

In our sample code, we will focus on language-tagged string literals. It is clear, however, that the described problems and proposed solutions also apply to language URIs for lexicons and lexical entries.

3 Finding solutions for N|uu and ||'Au, and Old French

We focus on varieties of N||ng and Old French to underline the fact that they are good examples of the need to preserve the languages and their historical stages as a key to understanding our cultural heritage: language is the storehouse of our culture, both past and present. It captures all aspects of life. It is subject to change and, thus, mirrors the development of our culture, of our state of mind, and of our social interaction through time.

[13] `https://iso639-3.sil.org/code/fro` [07-01-2019].

[14] Scripta is the term for the written form of a spoken dialect. Anglo-Norman is one of the varieties of Old French; it was spoken in England during the Anglo-Norman period.

[15] `https://www.iso.org/standard/43380.html` [07-01-2019].

[16] `https://iso639-3.sil.org/code/aoc`, `https://www.ethnologue.com/language/aoc` [11-01-2019].

[17] `https://iso639-3.sil.org/code/pro` [10-01-2019].

[18] `https://www.w3.org/TR/rdf11-concepts/#section-Graph-Literal`, `https://tools.ietf.org/html/bcp47#section-2.2.9` [11-01-2019]. – Note also that '@arab' is used to represent Arabic, although the ISO 639 code is 'ara', `https://iso639-3.sil.org/code/ara` [11-01-2019].

As little connected as N|uu, ||'Au, and Old French might ostensibly seem, they serve well to illustrate the problem: ISO 639 codes do not exist for N|uu, ||'Au, and the varieties of Old French. The atypicality of these languages highlights the relevance of the problem on a broader scale: more under-resourced, extinct or historical languages that are (currently) not included in the ISO 639 language code list will be published as LLOD.

3.1 N|uu and ||'Au

N||ng is the name of a dialect cluster of the !Ui-Tuu language family (formerly referred to as Southern Khoisan), spoken over a geographically large area in the southern Kalahari Desert; N|uu is the Western variety of N||ng, and ||'Au, the Eastern variety ([16, 11-17]; [33]; [5, 27]). Both dialects are near-extinct with two speakers for ||'Au and three speakers for N|uu as of 2013 (with the most fluent speaker of N|uu acting as a language teacher to young people); all N||ng speakers use Afrikaans as their main language [5, 15-16]. Since the late 19th Century, linguists have collected data of Khoisan[19] languages: this data is sparse, heterogeneous and difficult to access with misclassified languages, inappropriate language names and insufficient metadata as examples of the challenges faced, in addition to the identity of diverse corpora in archival material hard to assess, both in relation to each other and to modern languages ([16, 5-8]; [5, 2]). To document the many Khoisan languages is a challenge and a desideratum at the same time: encoding data following the Linked Data paradigm will convert the data into a valuable resource, possibly giving way to linguistic reconstructions using computational methods, where standard linguistic methodologies have been unable to yield meaningful results [5, 1]. Making accessible and preserving this data will contribute significantly to the exploration of the cultural heritage of mankind, with the collective group of Khoisan speakers being one of the few remaining hunter-gatherer cultures worldwide and the oldest existing human group today, according to genetic studies [29, 379].

3.1.1 Existing language codes

In order to convert the linguistic data of N|uu and ||'Au resources, we need an appropriate means to denote the languages in an unambiguous way, i.e., language codes to label the modeled elements of the linguistic resources in RDF. A language code for N||ng exists, i.e., ISO 639-3 'ngh'; this code is shared by both sub-languages N|uu and ||'Au.[20] However, according to the archival Khoisan 'doculects' discussed by [16, 16], the differences between the two language varieties are significant and, thus, explicit language codes for both ||'Au and N|uu are required.

Within MultiTree, a library of language relationships hosted by *The Linguist List*, the codes for N|uu and ||'Au are 'ngh-nuu' and 'ngh-aun' respectively.[21] Both are documented for 'Private Use', however their syntax does not meet the requirement defined by IETF's BCP 47, where the private use portion of the tag must be prepended with '`x-`' ([21]; [28, 4]). Furthermore, both the latter portions of MultiTree's codes, namely 'nuu' and 'aun', are pre-existing language codes, i.e., the former for the language Ngbundu (a language of the Congo area), and the latter for Molmo One (Papua New Guinea).[22] Despite the fact that

[19] The modern Khoisan languages are classified into three families and two isolates: the families Kx'a, !Ui-Tuu and Khoeid, and the isolates Hadza and Sandawe [5, 2].

[20] `https://iso639-3.sil.org/code/ngh` [29-12-2018].

[21] `http://www.multitree.org/codes/ngh-nuu`, `.../codes/ngh-aun` [20-06-2018].

[22] `https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry` [29-12-2018].

the use of *privateuse* sub-tags is by definition by private agreement only (*cf.* Point 2.2.7.5 of BCP 47, [28, 18]), it is clear that the use of MultiTree's language tags 'ngh-nuu' and 'ngh-aun' may lead to inadvertent misinterpretation when included in a language tag.

For this reason, we consider the use of Glottolog, a comprehensive catalogue of the world's lesser-known languages maintained by the Max Planck Institute for the Science of Human History. Their catalogue «assigns a unique and stable identifier (the Glottocode) to (in principle) all languoids, i.e. all families, languages, and dialects», [17]. Glottolog registers the two languages N|u and ||'Au (as sub languages of N||ng)[23] with the codes 'nuuu1242' and 'auni1243', respectively. However, as BCP 47 only allows for ISO 639 language codes in its *language* sub-tag, Glottolog is not recognized as a standard.

3.1.2 The use of *privateuse* sub-tag

In light of unambiguous language codes being available for the two Khoisan varieties, we propose to combine the ISO 639-3 code for the parent language N||ng, i.e., 'ngh', with the *privateuse* sub-tag '`x-`' and the respective Glottocodes stated above.

The language tags for N|uu and ||'Au can then be defined accordingly:

- N|uu: `ngh-x-nuuu1242`
- ||'Au: `ngh-x-auni1243`

A lexical concept, which can be linked to one or more senses in lexical entries from different languages, can be modeled as follows:

```
1  @PREFIX skos: <http://www.w3.org/2004/02/skos/core#> .
2  @PREFIX dbr:  <http://dbpedia.org/resource/> .
3  ...
4
5  :concept/000000001 a skos:Concept , ontolex:LexicalConcept ;
6     skos:example              "The belly is fat"@en ;
7     skos:example              "||'â he !qhûia."@ngh-x-nuuu1242 ;
8     ontolex:lexicalizedSense :en-n-belly#sense1 ;
9     ontolex:lexicalizedSense :ngh_x_nuuu1242-n-xa_belly#sense2 ;
10    ontolex:isConceptOf       dbr:Abdomen .
```

Where:

- Lines 6-7 show language-tagged strings, and line 7 the compiled language tag for N|uu.

3.2 Old French

Old French is the French spoken in the Middle Ages, and it can be more precisely defined as the umbrella term for the different Old French dialects[24] spoken in what is now France, parts of Belgium, England, Italy and the Holy Land. Its written resources date from 842 AD until c. 1350 AD (the border with Middle French) and its remarkable written tradition[25] serves to document its role as the most important vernacular of this time in Europe.

[23] `http://glottolog.org/resource/languoid/id/nuuu1241` [24-06-2018].

[24] The DEAF registers 30 varieties of Old French, Franco-Italian (a written, artificial language in the Middle Ages), and Judeo-French (sociolect), see Table 3, Appendix.

[25] Approx. 3,000 primary text sources transmitted within more than 10,000 manuscripts are registered by the *Complément bibliographique* of the DEAF, `http://www.deaf-page.de/bibl_neu.php` [07-01-2019].

3.2.1 Existing language codes

BCP 47's language tag offers a *variant* sub-tag that can be «used to indicate additional, well-recognized variations that define a language or its dialects that are not covered by other available subtags», where one or more variants can be used to form a language tag. Each of these variant sub-tags must be registered with IANA before use [28, 15]. Middle French is registered (ISO 639-3 code 'frm')[26] but no variants have been registered for Old French. IANA has registered Anglo-Norman (ISO 639-3 code 'xno'), but not as a sub-category of Old French, although it should be considered as such; the same applies to Zarphatic ('zrp': Judeo-French, spoken in the Middle Ages).

MultiTree lists Old French ('fro') and also the following child languages: Picard (ISO 639-3 code 'pcd'), Walloon ('wln'), and Zarphatic ('zrp'); Anglo-Norman ('xno') is not registered as a child language.[27] Although Walloon is registered as a child language of Old French, it is described as a living language; the same applies to Picard. Middle French is also registered as a child language of Old French, thus, following this logic, so should modern French. The hierarchization of Judeo-French (variety of Old French: sociolect) on the same level as Middle French (successor of Old French) and Picard / Walloon (modern dialects of the Picardy and Wallonia, respectively) conflates synchronic, diachronic, and geographical aspects.

Glottolog has assigned the identifier 'oldf1239' to Old French[28] but Glottolog does not register dialects of the medieval time period.[29] In addition to this flaw, Glottolog does not seem appropriate for the needs of linguists modeling data from the Romance languages, particularly with regard to old language stages. A closer look at Glottolog reveals major shortcomings in both the registration and the hierarchization of the Romance languages. E.g., Glottolog conflates diachronic and dialectal criteria within its hierarchies in several ways: Old French is registered (as a sub-entity of 'Oïl'[30]) at the same level as modern 'Central Oïl', Francoprovençalic (Romance language spoken in Eastern France), and Walloon. Following the hierarchy into the branches and sub-branches of 'Central Oïl' we find → Macro-French → Global French → French → a number of modern French dialects, but, also, Middle French and Anglo-Norman.[31] We deem necessary a thorough revision of the hierarchies, (re-)assembling both the dialects and regional varieties of modern French, and the historic stages of French.

3.2.2 Preliminary findings

The evaluation of language tags and language hierarchies in ISO 639, BCP 47, IANA, MultiTree, and Glottolog shows that the assignation of language codes to Old French dialects is not straightforward. At least for Anglo-Norman and Zarphatic, which we consider sub-categories of Old French, ISO 639-3 provides codes, i.e., 'xno' and 'zrp' respectively. These codes can be used for modeling lexemes and their graphical variants characterized as Anglo-Norman or Zarphatic. The following example for the Anglo-Norman noun *firbote*[32] illustrates this:

[26] The sub-tag 'frm-1606nict', `ftp://www.iana.org/assignments/lang-subtags-templates/1606nict.txt` [08-01-2019], does not depict a regional variety but the language documented by Jean Nicot in his *Thresor de la langue françoyse, tant ancienne que moderne*, Paris, from 1606.

[27] `http://www.multitree.org/codes/fro.html`; `.../pcd`; `.../wln`; `.../zrp`; `.../xno` [07-01-2019].

[28] `https://glottolog.org/resource/languoid/id/oldf1239` [07-01-2019].

[29] Old French is not available in the language collection of Ethnologue, as «ancient, classical, and long-extinct languages are not listed», `https://www.ethnologue.com/about/this-edition[29-12-2018]`.

[30] The term for the Romance varieties using an adaptation of the Vulgar Latin term *hoc ille* "this (is) it" as 'Yes'.

[31] More modern French dialects are found scattered in other sub-branches.

[32] Juridical term (in England) designating the right to take firewood from the land of a landlord, DEAF F 492,29, `https://deaf-server.adw.uni-heidelberg.de/lemme/firbote` [08-01-2019].

```
<firbote> a ontolex:LexicalEntry , ontolex:Word ;
  lexinfo:PartOfSpeech    lexinfo:Noun ;
  ontolex:canonicalForm   <firbote#form> .

<firbote#form> a ontolex:Form ;
  ontolex:writtenRep      "firbote"@xno .
```

3.2.3 The use of *privateuse* sub-tag

For the other Old French dialects and language varieties (see Table 3, Appendix), as language codes are not available, we again have to consider the use of BCP 47's *privateuse* sub-tag. E.g., a tag for the Old French variety spoken in Lorraine, a region in north-eastern France, could be defined as `fro-x-lorraine`. A simple example of an Old French word form characteristic of the Lorraine scripta is *feyvre*, a graphical variant of Old French *fevre* m.[33] This can be modeled as follows:

```
<fevre> a ontolex:LexicalEntry , ontolex:Word ;
   ontolex:canonicalForm <fevre#form_1> ;
   ontolex:otherForm     <fevre#form_2> .

# Old French standard form (lemma)
<fevre#form_1> a ontolex:Form ;
   ontolex:writtenRep    "fevre"@fro .

# graphical variant
<fevre#form_2> a ontolex:Form ;
   ontolex:writtenRep    "feyvre"@fro-x-lorraine .
```

3.2.4 Adding geographic information

The language tag can be further enriched by including geographic information, in line with established standards. There are several options available to us: (1) we could refer to the administrative region of France, (2) to the French *département*, or (3) use geographic coordinates. Both the administrative region and the *département* can be identified using the codes of the ISO 3166 standard for the administrative subdivisions of France.[34]

3.2.4.1 Administrative region and *département*

The area 'Lorraine' is part of the region Grand-Est (covering Alsace, Champagne, Ardenne, and Lorraine), thus the language tag can be defined as `fro-x-lorraine-FR-GES`.[35] However, the administrative region covers an area considerably larger than the geographic area of Lorraine, and thus does not map the area in question in a satisfying way. Another option would be to enrich the language tag by referring to the *département*, which would allow us to map the area more precisely.

Regarding options (1) and (2), the following concerns are raised:

[33] The smith, DEAF F 342,21, `https://deaf-server.adw.uni-heidelberg.de/lemme/fevre` [08-01-2019].
[34] `https://www.iso.org/obp/ui/#iso:code:3166:FR` [07-01-2019].
[35] *Ibid.*

(i) The administration of regions and *départements* is subject to change. As a consequence, the ISO 3166 codes are unstable, as evidenced by sub-divisions being allocated to new metropolitan regions in France as recently as 2016.[36]

(ii) The area in which an Old French dialect was spoken can embrace several modern regions, e.g., 'Nord-Est' and 'Sud-Ouest' (see Table 3, Appendix), or *départements*: e.g., contemporary Lorraine consists of not one but four *départements*, i.e., Meurthe-et-Moselle (ISO 3166-2:FR-54), Meuse (ISO 3166-2:FR-55), Moselle (ISO 3166-2:FR-57), and Vosges (ISO 3166-2:FR-88); the historical region also comprises the contemporary *département* Haute-Marne (ISO 3166-2:FR-52). As a result, either more than one region may need to be included in the sub-tag, indicating (imprecisely) the geographical boundary in which the dialect was spoken, or the RDF triples must be manifolded: when modeling a lexeme or a graphical variant of a lexeme characterized as Lorraine, e.g., within the data of the DEAF dictionary, the inclusion of the codes for the *départements* into the language tag requires duplicating the RDF triples, thus creating somewhat unwieldy data.

(iii) The boundaries of the regions are modern-day boundaries which may not necessarily align to the boundaries of a previous time. This leads to a dissatisfying mapping of said area.

3.2.4.2 Geographic coordinates

As a third option, we consider the inclusion of geographic coordinates in the language tag. To do this, we map the (approximate) geographic distribution of 'Lorraine' to coordinates, assuming that the last coordinate is the same as the first coordinate, and the coordinates are ordered in a counterclockwise direction, thus creating a polygon shape [25]. Each coordinate can be compressed using Geohash, a system for encoding geographic coordinates into a base32 string, which would also format each latitude and longitude value in a syntax acceptable for BCP 47.[37] As precision down to the nearest meter is not necessary, the Geohash length could be limited to five characters,[38] rendering the coordinate in an approximate area that is $\leq$ 4.89 x 4.89 kilometers.[39]

As using the geographic coordinates to map the modern-day distribution of 'Lorraine' would lead to the same dissatisfying result (*cf.* 3.2.4.1), we draw on a map of the *Französisches Etymologisches Wörterbuch* – FEW [34] that includes historical information, see Fig. 1.[40]

To derive the geographic coordinates for the old dialect of 'Lorraine', we take this map as a substratum and the result is the following:
(4.91473,49.62686), (4.6696405,48.0428789), (5.59192,47.6435), (6.858446002006532, 47.883257283545234), (7.2386756,48.4086571), (5.81263,49.72584), (4.91473,49.62686)[41]

[36] `https://www.iso.org/obp/ui/#iso:code:3166:FR` [29-12-2018].

[37] Geohash 2018, `https://en.wikipedia.org/wiki/Geohash` [31-12-2018]; Geo-shape datatype 2018, `https://www.elastic.co/guide/en/elasticsearch/reference/current/geo-shape.html` [31-12-2018].

[38] Or less, depending on the extent of the geographical distribution of the dialect being mapped.

[39] `https://www.movable-type.co.uk/scripts/geohash.html` [31-12-2018].

[40] In the possession of the editorial office of the DEAF is a 40-year-old, battered copy of the map of France that is included in the *Beiheft* of the FEW. This copy contains the boundaries of the areas where the Old French dialects were spoken, sketched in by hand (in yellow) by Frankwalt Möhren, co-founder of the DEAF (and, also, valuable notes and comments, e.g., the indication 'Orval: Bier!': the Abbey of Orval in Villers-devant-Orval is the home of the famous top-fermented beer 'Orval').

[41] Ordering is latitude then longitude.

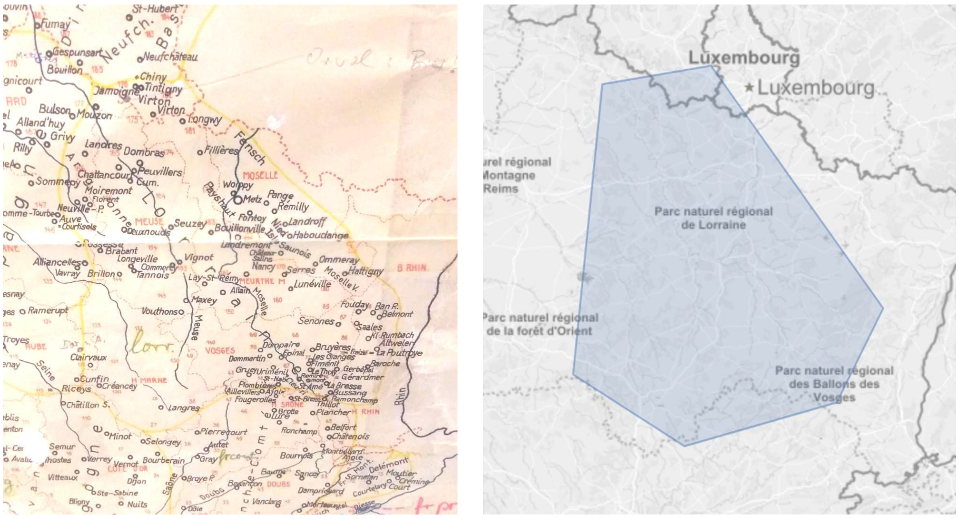
Luxembourg
Luxembourg
Parc naturel régional de Lorraine
Parc naturel régional de la forêt d'Orient
Parc naturel régional des Ballons des Vosges
urel régional
Montagne
Reims

Figure 1 'Old Lorraine' area: Extract of the map of the FEW (left), mapped using geographic coordinates (right).

Each longitude and latitude coordinate can be converted to a Geohash, to a precision of five characters: `t0g7c, t0f4t, t0czu, t14p1, t163j, t1535, t0g7c`.

As the last coordinate is the same as the first coordinate, the last one can be excluded, and as only alphanumeric characters and hyphens are allowed by BCP 47, every Geohash, with the exception of the first one, is prepended with '`--`' to serve as an internal delimiter; the language code for the language, dialect and region can thus be presented as follows: `fro-x-lorraine-t0g7c--t0f4t--t0czu--t14p1--t163j--t1535`.

The use of a historical map as a source of information to enrich a language tag with geographic coordinates, as demonstrated for medieval Lorraine, seems very promising to us regarding our aim: the unambiguous and historically-correct tagging of languages.

A further possibility is to include the period of time within the language tag, e.g., `fro-x-lorraine-t0g7c--t0f4t--t0czu--t14p1--t163j--t1535-850AD--1350AD`, where `850AD--1350AD` depicts the time range.[42]

BCP 47 specifies the maximum length of a sub-tag to be of eight characters (+ two for '`x-`', see [28, 6]). However, numerous examples of the *privateuse* sub-tags exceed this maximum length [28, 56,81]. Thus, we conclude that there is not an upper limit to the length of the *privateuse* sub-tag, except that pertaining to buffer overflow [28, 63,71-72].

4 Discussion

The examples, N|uu and ‖'Au, and Old French, demonstrate that there is not a single, encompassing solution that can be applied to all languages. For each of the three languages, a custom approach, in conjunction with the *privateuse* sub-tag from BCP 47's language tag, has had to be adopted. However, with each example, a tentative pattern for the *privateuse* sub-tag has emerged: each part within the *privateuse* sub-tag can be assigned to a category, as listed in Table 1, and the *privateuse* sub-tag can consist of one or more parts.

[42] In the case of Old French, this seems dispensable since the code 'fro' contains this information, however it could be valuable when identifying a language where the geographical distribution changes significantly.

Table 1 The categorization of parts in a *privateuse* sub-tag.

Part	**Description**
language	A language, dialect or pidgin not in ISO 639
otherlect	An ethnolect, sociolect, or idiolect
timeperiod	If not modern-day; not equivalent to the time period specified by the language code
region	A geographic, politic or administrative region

Using the categories identified in Table 1, we thus propose the following pattern for the *privateuse* sub-tag of a language tag, with each part separated by a ‘-’:

`x-language-otherlect-timeperiod-region`

Within BCP 47, the format of the language tag has been designed such that each sub-tag can be identified on the basis of its length, position in the tag, and its content, and each sub-tag is typically a code from an ISO standard or registry [28, 8]. However, this requirement can be limiting and inflexible. In order to identify each part in the *privateuse* sub-tag pattern, we propose prepending each part with a key consisting of 2 digits, from 0 - 9, with the first digit, Key 1, indicating the category, and the second digit, Key 2, indicating the content in relation to Key 1, as shown in Table 2. This way, each part can be of variable length, thus allowing for greater flexibility. For example, a part that is categorized as *language* can be prepended with ‘10’, where ‘1’ indicates that it is *language* and ‘0’ indicates that the language is user-defined information. The tags can, thus, be rewritten as follows:

- N|uu dialect: `ngh-x-01nuuu1242`
- ‖’Au dialect: `ngh-x-01auni1243`
- Old French, Lorraine dialect:
  `fro-x-00lorraine-30t0g7c--t0f4t--t0czu--t14p1--t163j--t1535`

Table 2 The key for each part of the *privateuse* sub-tag.

Part	**Key 1**	**Key 2**
language	0	0 = User-defined 1 = Glottocode
otherlect	1	0 = User-defined 1 = Glottocode
timeperiod	2	0 = one year only, BC 1 = one year only, AD 2 = start:BC - end:BC 3 = start:BC - end:AD 4 = start:AD - end:AD
region	3	0 = Geohashed latitude and longitude coordinates – polygon 1 = Geohashed latitude and longitude coordinates – point only 2 = URI to GeoJSON-LD 3 = Code from ISO 3166

The interpretation of a language tag which contains multiple sub-tags can be obscure and requires human inspection. By (1) categorizing the *privateuse* sub-tag into parts, then (2) defining a key for each part, and (3) defining rules for each key, it not only allows for more accurate interpretation, by both human and machine, but it can also lead to increased shared agreement for a compiled language tag.

5 Conclusion and Future Work

In this paper, we have discussed the shortcomings of language tags in the context of modeling data from lesser-known languages as LD. For two under-resourced language varieties and one historical language stage we have proposed solutions using the *privateuse* sub-tag, with the addition of geographic information. This can improve a language tag so that it reflects the diachronic, synchronic and dialectal aspects of the language in question.

The proposed rule-based pattern for the *privateuse* sub-tag is not intended to be used in place of other sub-tags in the language tag, nor is it intended to replace the work of existing standards and bodies. The W3C Internationalization (i18n) Interest Group[43] serves to connect a large group of people on the topic of internationalization on the Web. The authors intend to contribute to the discussions of the group, submitting the proposals outlined in this paper for further feedback. Also, the authors and C. Maria Keet propose MoLA, a **Mo**del for **L**anguage **A**nnotation (`https://ontology.londisizwe.org/mola`) [14]. MoLA has been developed to provide a vocabulary for language annotation in RDF, which enables custom language tags to be defined, and for said language tags to be associated with both a time period and region.

Defining a pattern for the *privateuse* sub-tag can lead to discussions which can improve the next iteration of BCP 47, as well as to increased interoperability within the context of LLOD so as to render language identification more accurate. This in turn can lead to shared agreement between lexical resources and to re-use, an important notion in a multilingual Semantic Web.

References

1 K. Baldinger. *Dictionnaire étymologique de l'ancien français – DEAF*. Presses de L'Université Laval / Niemeyer / De Gruyter, Québec/Tübingen/Berlin, since 1971. [Continued by Frankwalt Möhren, and Thomas Städtler; DEAF*él*: `https://deaf-server.adw.uni-heidelberg.de`].

2 A. Bellandi, E. Giovannetti, and A. Weingart. Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information*, 9 (3), 52, 2018.

3 T. Berners-Lee. *Linked Data*. World Wide Web Consortium, 2006.

4 Ch. Bizer, T. Heath, and T. Berners-Lee. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5:1–22, 2009.

5 M. Brenzinger. The twelve modern Khoisan languages. In A. Witzlack-Makarevich and M. Ernszt, editors, *Khoisan Languages and Linguistics: Proceedings of the 3rd International Symposium July 6-10, 2008, Riezlern / Kleinwalsertal*, pages 1–32. Köppe Verlag, 2008.

6 Ch. Chiarcos, J. McCrae, Ph. Cimiano, and Ch. Fellbaum. Towards Open Data for Linguistics: Lexical Linked Data. In A. Oltramari et al., editor, *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, pages 7–25. Springer, Berlin, Heidelberg, 2013.

7 Ch. Chiarcos and M. Sukhareva. Linking Etymological Databases. A Case Study in Germanic. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page 41, 2014.

8 P. Cimiano, J.P. McCrae, and P. Buitelaar. Lexicon model for ontologies: community report, 10 May 2016. Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA), 2016. URL: `https://www.w3.org/2016/05/ontolex/`.

9 R. Cyganiak, D. Wood, and M. Lanthaler. RDF 1.1. concepts and abstract syntax: W3C recommendation 25 February 2014, 2014. URL: `https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/`.

[43] `https://www.w3.org/International/ig/Overview` [15-03-2019].

10 Gerard de Melo. Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web*, 6(4):393–400, August 2015.

11 Th. Declerck, E. Wandl-Vogt, and K. Mörth. Towards a Pan European Lexicography by Means of Linked (Open) Data. In I. Kosem et. al., editor, *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, pages 342–355. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 2015.

12 International Organization for Standardization. Language codes – ISO 639. URL: `https://www.iso.org/iso-639-language-codes.html`.

13 F. Gillis-Webber. Conversion of the English-Xhosa Dictionary for Nurses to a linguistic linked data framework. *Information*, 9(11), 2018. `doi:10.3390/info9110274`.

14 F. Gillis-Webber, S. Tittel, and C. M. Keet. A Model for Language Annotations on the Web, 2019. (submitted).

15 J. Gracia, M. Villegas, A. Gómez-Pérez, and N. Bel. The Apertium Bilingual Dictionaries on the Web of Data. In *Semantic Web – Interoperability, Usability, Applicability*, pages 1–10. IOS Press, 2017.

16 R. Güldermann. Towards casting a wider net over N‖ng: chances and challenges of archival Khoisan resources, 2014. URL: `https://www.iaaw.hu-berlin.de/de/region/afrika/afrika/linguistik/mitarbeiter/1683070/dokumente/2014-03-cape-town-nng-h`.

17 H. Hammarström, R. Forkel, and M. Haspelmath. Glottolog 3.3., 2018. accesssed 21-02-2019.

18 SIL International. ISO 639-3: Relationship between ISO 639-3 and the other parts of ISO 639, 2017. URL: `https://iso639-3.sil.org/about/relationships`.

19 SIL International. ISO 639-3: Scope of denotation for language identifiers, 2017. URL: `https://iso639-3.sil.org/about/scope`.

20 SIL International. ISO 639-3: Types of individual languages, 2017. URL: `https://iso639-3.sil.org/about/types`.

21 R. Ishida. Language Tags in HTML and XML, 2014. URL: `https://www.w3.org/International/articles/language-tags/index.en`.

22 F. Khan, J.E. Díaz-Vera, and M. Monachini. The Representation of an Old English Emotion Lexicon as Linked Open Data. In John P. McCrae et al., editor, *Proceedings of the LREC 2016 Workshop "LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources", 24 May 2016 – Portorož, Slovenia*, pages 73–76, 2016.

23 G. Köbler. *Wörterbuch des althochdeutschen Sprachschatzes.* Schöningh, Paderborn, 1993.

24 L. Lezcano, S. Sánchez-Alonso, and A. Roa-Valverde. A Survey on the Exchange of Linguistic Resources. *Program*, 47,3:263–281, 2013.

25 J. Lieberman, R. Singh, and Ch. Goad. W3C geospatial vocabulary: W3C incubator group report 23 October 2007, 2007.

26 J.P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, and P. Cimiano. The OntoLex-Lemon model: Development and Applications. In *Proceedings of ELEX 2017: Lexicography from Scratch. September 2017*, pages 19–21, 2017.

27 S. Moran and M. Brümmer. Lemon-aid: Using Lemon to Aid Quantitative Historical Linguistic Analysis. In Ch. Chiarcos et al., editor, *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013), Pisa, September 2013*, pages 28–33. Ass. for Comp. Linguistics, 2013.

28 A. Phillips and M. Davis. Tags for Identifiying Languages. *BCP*, 47, 2009.

29 C.M. Schlebusch, P. Skoglund, and P. Sjödin et al. Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science*, 338(6105):374–379, 2012.

30 S. Shah and M. Brenzinger. *Ouma Geelmeid ke kx'u ‖xa‖xa N|uu.* Centre for African Language Diversity, University of Cape Town, Cape Town, 2016.

31 S. Tittel, H. Bermúdez-Sabel, and Ch. Chiarcos. Using RDFa to Link Text and Dictionary Data for Medieval French. In J.P. McCrae et al., editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 6th Workshop on Linked Data in Linguistics (LDL-2018), Miyazaki, Japan, 2018*, pages 30–38, Paris (ELRA), 2018.

32 S. Tittel and Ch. Chiarcos. Historical Lexicography of Old French and Linked Open Data: Transforming the Resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018), Miyazaki, Japan, 2018*, pages 58–66, Paris (ELRA), 2018.

33 M. Van Der Merwe. Giving breath to a dying history, 2015. URL: `https://www.dailymaverick.co.za/article/2015-01-23-giving-breath-to-a-dying-history/#.Wyvou9WFMsk`.

34 W. von Wartburg. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes – FEW.* ATILF, since 1922. [Continued by O. Jänicke, C.T. Gossen, J.-P. Chambon, J.-P. Chauveau, and Yan Greub].

35 D. Wood, M. Zaidman, L. Ruth, and M. Hausenblas. *Linked data: structured data on the web.* Manning Publications Co., New York, 2014.

A Old French dialects

Table 3 List of Old French dialects (described in French) registered by the DEAF.

Abbrev.	Language	Abbrev.	Language
afr.	ancien français	saint.	saintongeais
mfr.	moyen français	tour.	tourangeau
fr. du 16^{e}s.	français du 16^e siècle	orl.	orléanais
fr.dial.	français dialectal	bourb.	bourbonnais
frc.	francien (français de l'Ile de France)	bourg.	bourguignon
pic.	picard	lyon.	lyonnais
flandr.	français de la Flandre française	frcomt.	franc-comtois
hain.	hennuyer	francoit.	franco-italien
art.	artésien	Nord-Est	
wall.	wallon	Nord	
liég.	liégeois	Nord-Ouest	
champ.	champenois	Ouest	
lorr.	lorrain	Sud-Ouest	
norm.	normand	Centre	
agn.	anglo-normand	Est	
hbret.	haut-breton	Sud-Est	
ang.	angevin	Terre Sainte	
poit.	poitevin	judéofr.	judéofrançais