Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3333–3339
Marseille, 11–16 May 2020

A Framework for Shared Agreement of Language Tags beyond ISO 639

Frances Gillis-Webber, Sabine Tittel
Department of Computer Science, University of Cape Town, Heidelberg Academy of Sciences and Humanities
Cape Town, South Africa, Heidelberg, Germany
fran@fynbosch.com, sabine.tittel@urz.uni-heidelberg.de

Abstract

The identification and annotation of languages in an unambiguous and standardized way is essential for the description of linguistic data. It is the prerequisite for machine-based interpretation, aggregation, and re-use of the data with respect to different languages. This makes it a key aspect especially for Linked Data and the multilingual Semantic Web. The standard for language tags is defined by IETF's BCP 47 and ISO 639 provides the language codes that are the tags' main constituents. However, for the identification of lesser-known languages, endangered languages, regional varieties or historical stages of a language, the ISO 639 codes are insufficient. Also, the optional language sub-tags compliant with BCP 47 do not offer a possibility fine-grained enough to represent linguistic variation. We propose a versatile pattern that extends the BCP 47 sub-tag *privateuse* and is, thus, able to overcome the limits of BCP 47 and ISO 639. Sufficient coverage of the pattern is demonstrated with the use case of linguistic Linked Data of the endangered Gascon language. We show how to use a URI shortcode for the extended sub-tag, making the length compliant with BCP 47. We achieve this with a web application and API developed to encode and decode the language tag.

Keywords: language tags, language codes, Linked Data

1. Introduction

Data modeling and the processing of linguistic resources requires the use of language tags. Language tags serve to identify the language of a resource and the linguistic entities therein. In particular, language tagging is a key aspect when modeling linguistic resources in *Resource Description Framework* (RDF) (Manola and Miller, 2014) following the principles of Linked Data (LD) (Berners-Lee, 2006).

Language tags are typically compiled of language codes, with additional sub-tags where applicable. By adhering to established standards, language tags support a machine-based, cross-resource interoperability with respect to the identification of the language of given data. BCP 47 (*Best Current Practice*, by the *Internet Engineering Task Force* (IETF)) defines the standard for language tags (Phillips and Davis, 2009). A language code, in the form of a two- or three-character identification such as 'en' for English, is the main constituent of a language tag and is provided by the ISO 639 standard (International Organization for Standardization, nd). The codes must be unambiguous to be able to disambiguate the case when one language name refers to several languages, or one language is denoted by several names. However, for the identification and annotation of lesser-known languages, endangered languages, regional varieties or historical stages of a language, the ISO 639 codes are insufficient. Also, the optional extended language sub-tags compliant with BCP 47 do not offer a possibility fine-grained enough to represent the world's modern and historic linguistic diversity.

In Gillis-Webber and Tittel (2019), we proposed a versatile pattern that extends the BCP 47 sub-tag *privateuse* and is, thus, able to overcome the limits of BCP 47 and ISO 639 (but still staying compliant with BCP 47). We demonstrated the pattern for (i) N|uu and ǁ'Au, two dialects from N‖ng, a critically endangered non-Bantu click language in Southern Africa, and (ii) Old French. In Tittel and Gillis-Webber (2019), we continued with the same pattern, this time applying it to diachronic and diatopic specifications of French (Old, Middle and Modern). We expanded on the pattern in a way that allowed for URI shortcodes, yet we did not cover this beyond a cursory mention. We also only briefly introduced a web application, to be used as a user-friendly way to generate and decode language tags using our pattern.

In this paper, the web application is our focus, as well as a newly-developed API to decode language tags. We also expand fully on the notion of URI shortcodes: We demonstrate how, when used in conjunction with the website and API, the URI shortcode ensures both compliancy with BCP 47 and allows for detailed information about that language tag, thus addressing any criticism regarding overly-long language tags.

Sufficient coverage of the pattern, the web application, the API and the URI shortcode is demonstrated using language data from Gascon, to be published as Linked Data. The Gascon language is an under-resourced, endangered European language for which an ISO 639 code is not available. It is spoken in the southwestern area of France by a shrinking minority (in particular in networks of mainly older people); the number of Gascon speakers in this region is around 500,000 in total and varies from 3% (city of Bordeaux) to 35% (rural areas) of the population (Moreux, 2004, 25). (In ISO 639, Gascon is included in the notion of 'Occitan'.[1])

To focus on the modeling of endangered languages with a limited number of linguistic resources is an important step to promote linguistic diversity. To survey and document linguistic diversity plays a crucial role for the understanding of different cultures and their often interacting historical developments: Languages are the storehouse of the past and the mirror and instrument of the socio-cultural interaction of a society. In this context, the cultural value of endan-

[1] With ISO 639-1 'oc' / 639-3 'oci', `https://iso639-3.sil.org/code/oci` [29-11-2019].

gered languages is becoming increasingly recognized. To raise awareness of endangered languages and their important contribution to the world's rich cultural diversity has been defined a major aim of the United Nations / UNESCO International Year of Indigenous Languages 2019.[2] This aim is supported by an enhanced emphasis on the scientific exploration and documentation of endangered languages.

The remainder of the paper is structured as follows: After a short introduction to language tagging and language codes in section 2., we introduce the extension pattern of language tags in section 3. In section 4., we describe the web application and API for generating and decoding language tags. We then demonstrate the pattern and the web application, using the Gascon language and its variants as a use case in section 5. Following this, we elaborate on the generation of URI shortcodes in section 6., still using Gascon as the use case. We conclude the paper in section 7. with an illustration of the application of our pattern using sample data from Old Gascon.

2. Language Tagging

A language tag is a means to identify a natural language. The use of language tags is a mandatory requirement for a number of computational standards. These include HTML (where the attribute `lang` in the top element, e.g., `<html lang="en">`, defines the language of the document and can also be used in any other HTML element), XML (using the attribute `xml:lang` for pages that are delivered in XML and any XML element therein), and HTTP (Fielding and Reschke, 2014, 3.1.3.1). The use of language tags is also obligatory when modeling linguistic data as Linked Data / RDF. There, the tag specifies the language of the linguistic resource and also of its elements, such as lexemes, morphemes, phonemes, and syntactic patterns.

IETF's BCP 47 Tags for Identifying Languages (Phillips and Davis, 2009) defines the standard for language tags as a sequence of one or more case-insensitive sub-tags separated by a hyphen character (Phillips and Davis, 2009, 2.1.1). A language tag typically begins with a standardized, unique language code defined by ISO 639 (with Parts 1-3 International Organization for Standardization (nd)) which can be followed by sub-tags refining or narrowing the range of the encoded language in the following form: `language-extlang-script-region-variant-extension-privateuse`.

BCP 47 specifies that these sub-tags are identified on the basis of their length, position in the tag, and their content and that they are recorded by an ISO standard or a registry (Phillips and Davis, 2009, 8).

The practice of language tagging is straightforward whenever ISO 639 provides a language code for a given language, i.e., for a large number of modern and well-known languages. However, this is neither the case for historical stages of well-resourced languages, nor for modern, under-resourced or minority languages including their historical language stages. ISO 639-3 lists 7,865 languages (International Organization for Standardization, nd), yet the number of today's languages spoken in the world is estimated up to 10,000, with some 150,000 extinct languages (Crystal, 2010, 294-295). It becomes clear that the established standards have limited coverage of the world's linguistic diversity, in particular when focusing on regional language variation, minority languages and diachronic language stages. We discuss the shortcomings of ISO 639 language codes and we also show that language catalogues like Glottolog, Ethnologue, and MultiTree[3] are not valid alternatives in Gillis-Webber and Tittel (2019), Tittel and Gillis-Webber (2019) and Gillis-Webber et al. (2019).

With the *privateuse* sub-tag, users can create their own language tag using the main parts of BCP 47. However, the use of the *privateuse* sub-tag is, by definition, by private agreement only (Phillips and Davis, 2009, 8). Thus, when the *privateuse* sub-tag is used, collaboration requires a channel of communication beyond the established standard so that the meaning of the tag can be identified, and the same tag can be reused by other users.

3. Pattern for the Extension of Language Tags

To offer adequate possibilities for the language tagging of non-represented languages or lects (language varieties) in ISO 639, Gillis-Webber and Tittel (2019) and Tittel and Gillis-Webber (2019) proposed a fine-grained pattern for extending the *privateuse* sub-tag in the following form: `x-language-otherlect-timeperiod-region-uri`, where `x-` is the BCP 47 requirement that indicates *privateuse*, and the pattern that follows is divided into 'language' (designating a language, dialect, patois or pidgin), 'otherlect' (ethnolect, sociolect, or idiolect), 'timeperiod', 'region', and URI. For each part in the *privateuse* sub-tag, Gillis-Webber and Tittel (2019, 552) define a system of two keys to identify each part respectively. This system of keys thus allows for flexibility of position and content, unlike that for other sub-tags in BCP 47.

Reproducing the table from Tittel and Gillis-Webber (2019), we extend it in Table 1 to include both Wikidata and VIAF identifiers, as well as two 'timeperiod' options. Wikidata is a knowledge base of structured data, intended for use by other 'Wiki-' projects (such as Wikipedia). It has a free licence and data dumps are regularly made available for download as JSON, RDF and XML.[4] Identifiers from the *Virtual International Authority File* (VIAF)[5] allows for regional specification. For the 'timeperiod' part of the pattern, it now allows for a year to start in BC/AD and continue up until present-day.

As Table 1 demonstrates, this pattern allows for an unambiguous and detailed annotation of a lect in time and space: (i) It facilitates a diachronic specification of a lect that only possesses an ISO 639 code for the modern stage of the language. (ii) It enables a geospatial specification of a lect that possesses an ISO 639 code for a broad and unspecific range. The specification can be noted with a precision from a stan-

[2] `https://en.iyil2019.org/` [25-11-2019].

[3] `https://glottolog.org/`, `www.ethnologue.com`, `http://multitree.org/` 30-11-2019].

[4] `https://www.wikidata.org/` [30-11-2019].

[5] `https://viaf.org/` [01-12-2019].

Part	Key 1	Key 2
language	0	0 = User-defined 1 = Glottocode 2 = Identifier from Wikidata
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 5 = start:BC to present day 6 = start:AD to present day
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 4 = Identifier from GeoNames 5 = Identifier from VIAF
URI	4	0 = URI shortcode from https://londisizwe.org/language-tags/

Table 1: The keys for each part in the *privateuse* tag.

dardized region (as in ISO 3166 [6]) to geohashed latitude and longitude coordinates for polygons and single points. Both types of specification can be combined in one language tag.

4. Web Application for Generating and Decoding Language Tags

A web application has been developed to assist an end-user with the generation and decoding of language tags. Although use of this application is not a requirement when compiling a language tag according to our pattern, it serves as an auxiliary tool.

The web application can be found at: `https://londisizwe.org/language-tags/`. A screenshot of the homepage is shown in Fig. 1.

The following activities are supported:

1. **Generating a language tag**
   - The user can complete the fields which apply to their language tag.
   - The language tag is generated in real-time and can be copied to the user's clipboard or saved to their profile (logging in is required for the latter action).
2. **Decoding a language tag**
   - The user can enter a language tag for decoding.

[6] `https://www.iso.org/obp/ui/#iso:code:3166:FR` [27-11-2019].

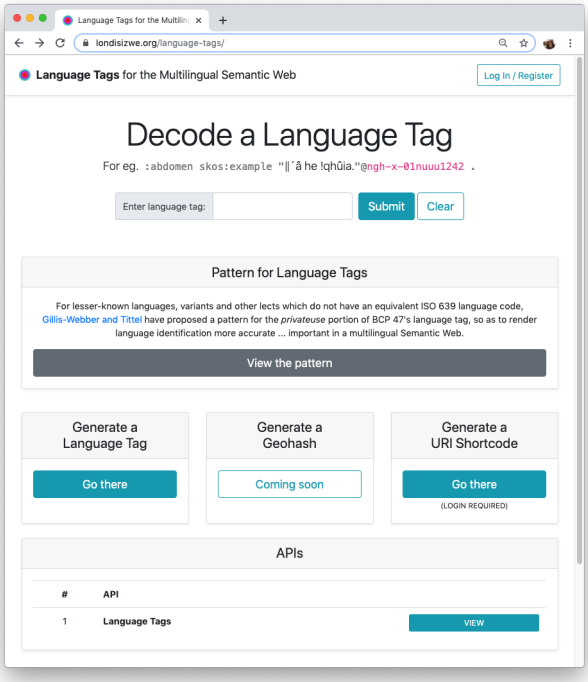

Figure 1: Web application 'Language Tags for the Multilingual Semantic Web' for generating and decoding language tags.

   - A language tag can be an ISO 639 language code or a language tag encoded using our pattern.
   - Results are returned in JSON format.
3. **Registration and Login**
   - The OAuth 2.0[7] protocol is used for authentication, with GitHub client credentials currently supported.
   - No passwords are stored and only publicly available information provided by GitHub for that user is requested.
   - Once successfully logged in, the user is presented with their dashboard. This shows their details, saved language tags, saved URI shortcodes, generated API keys, and APIs to which they have subscribed.
4. **Using the Language Tags API**
   - This is a REST API[8] with stateless operations.
   - To make use of this API, the user must first subscribe to it (logging in is required for this action). Thereafter, an API key should be manually generated.
   - The *decode* method can be called using GET. In the request body, the language tag should be included as a parameter. The response will be the same as that from Point 2.

In the section that follows, the generation of language tags

[7] `https://oauth.net/2/` [01-12-2019].

[8] `https://en.wikipedia.org/wiki/Representational_state_transfer` [01-12-2019].

for Gascon and one of its varieties, i.e., Béarnais, is demonstrated, using the web application.

5. Generating Language Tags for Gascon

Gascon is a language spoken in the southwestern part of France (see Fig. 2) by around 500,000 people (as of 2004[9]), among them only few people of young age (Moreux, 2004, 25-26; 44). French is the sole official language, but other languages and their regional varieties are also spoken in France, e.g., Breton (a Celtic variety spoken in Brittany), the Franco-Provençal language (a Romance language spoken in East-central France), Occitan (Romance language spoken in Southern France) and Gascon. Gascon has been one of the Romance languages spoken in what is today France from the Middle Ages on while French was only introduced to the Gascon speaking region in the 17th century (Glessgen et al., 2008, 23,1,1780b).

Since then, its prestige and use, both as a written and spoken language, has declined as a result of the process of Francization, whose aim was to strengthen the dominance of French and its role as the pillar of a French nation, in particular after the French Revolution (Abalain, 2007; Rey et al., 2007, 120).[10] It is widely discussed whether Gascon is to be considered a Romance language on its own or a variety of Occitan (Field, 2009). Regardless of this discussion, it is a recognized fact that Gascon belongs to the endangered languages in Europe (Moseley, 2010).

Gascon is spoken in Gascony and Béarn, roughly within the triangle bordered by the Garonne river, the Atlantic Ocean and the border to Spain. The language area covers (parts of) the French *départements* Pyrénées-Atlantiques, Hautes-Pyrénées, Landes, Gers, Gironde, Lot-et-Garonne, Haute-Garonne, and Ariège and, also, the Aran Valley of Catalonia, Spain. Gascon can be divided into several regional varieties or lects, e.g., Béarnais, Landais and Armagnacais.

To model linguistic data of Gascon and its varieties as, for example, Linked Data, language tags are required to clearly identify the lects. However, the ISO 639 standard only offers a code for Occitan, i.e., ISO 639-1 'oc' ('Occitan (post 1500)') and ISO 639-3 'oci' respectively.[11] A Gascon code

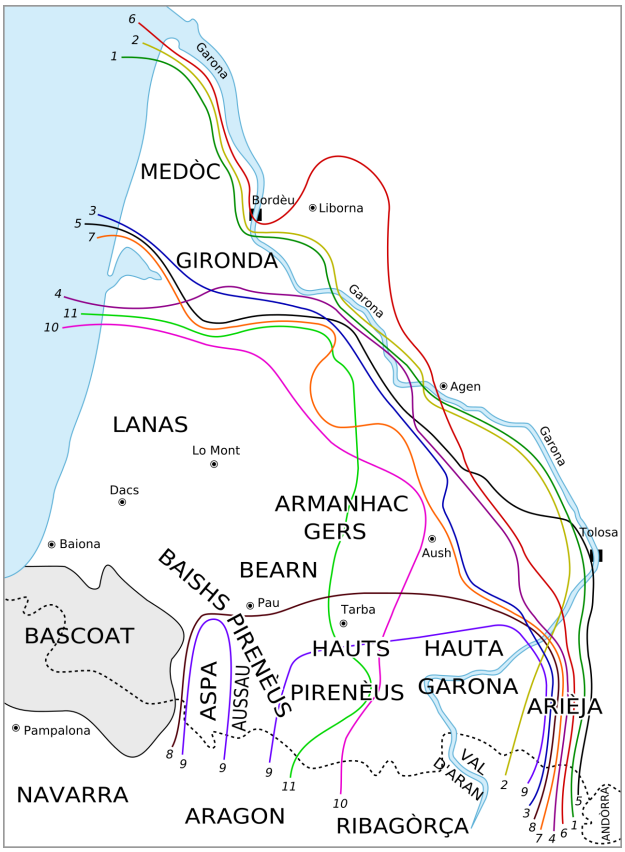

Figure 2: The Gascon speaking area shown with the isoglosses that delimit or intersect the area (CC BY-SA 3.0).

'gsc' did previously exist in ISO 639-3 but its status is deprecated and it was merged into 'oci' in March 2007.[12] This ignores the long debate whether Gascon is a language distinct from Occitan or a dialect of it. Furthermore, there are no ISO 639-codes for varieties like Landais spoken in the *département* Landes, nor for (the several varieties) of Béarnais (Béarn) that is considerably distinct from the other lects spoken in Gascony (Moreux, 2004).

Glottolog does register Gascon and varieties: Glottolog's code is 'gasc1240'[13], but, as with the language variation within French (cp. Gillis-Webber and Tittel (2019, 4:8)), the hierarchy of the lects subordinated to Gascon is flawed. In a nutshell, the situation hampers a fine-grained representation of the linguistic diversity in southern France.

Using Modern and Old Gascon and one of its varieties, i.e Modern and Old Béarnais, as a use case, we have identified the following language tags for each, in accordance with the pattern from Table 1:

1. **Modern Gascon:**
   - `oc-x-02q35735`
   - For `02q35735`, the '02' key indicates that it is an identifier from Wikidata for the Gascon language.

[9] Other sources give the number of 250,000 in 1990 (Field, 2009, 756).

[10] Since 2008, the newly added article 75-1 of the French Constitution officially recognizes these regional languages spoken in France as belonging to the patrimony of France ("Les langues régionales appartiennent au patrimoine de la France", see `https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006071194#75-1` [25-11-2019]). This is a potential step towards a future ratification by France of the European treaty 'European Charter for Regional or Minority Languages' (ECRML) whose aim is to protect and promote historical regional and minority languages in Europe (see "The European Charter for Regional or Minority Languages is the European convention for the protection and promotion of languages used by traditional minorities", `https://www.coe.int/en/web/european-charter-regional-or-minority-languages` [23-11-2019].

[11] `https://iso639-3.sil.org/code/oci` [25-11-2019].

[12] `https://iso639-3.sil.org/code/gsc` [25-11-2019].

[13] `https://glottolog.org/resource/languoid/id/gasc1240` [25-11-2019].

2. **Old Gascon:**
   - `oc-x-02q35735-241050--1500`
   - For `241050--1500`, the '24' key indicates that there is a time period applicable to the language tag.
     The time period is from 1050AD to 1500AD. The determination of the date range of Old Gascon and its varieties is barely discussed in the literature. The oldest evidences of a Gascon scripta (written evidence of the spoken language) date from the 11th century, attested by the source material of the *Dictionnaire onomasiologique de l'ancien gascon* (DAG, Baldinger (since 1975))[14], the Proto-Gascon existing since before 600 A.D. (Chambon and Greub, 2002, 482). As an end date for Old Gascon (and Old Béarnais respectively) we can determine 'ca. 1500', in line with the DAG.[15]

3. **Modern Béarnais:**
   - `oc-x-00bearnais-35136075732`
   - For `00bearnais`, the '00' key indicates that this is a user-defined identifier for a language; for `35136075732`, the '35' indicates that the region corresponds to the historical province of Béarn identified in VIAF: `http://viaf.org/viaf/136075732` [01-12-2019]. To model the area Old Béarnais was spoken (and written) in a more accurate way (considering in detail the medieval textual resources of the area), a polygon with Geohashed latitude and longitude coordinates can be used (with key '30'; examples are shown in Tittel and Gillis-Webber (2019, 558-559)).

4. **Old Béarnais:**
   - `oc-x-00bearnais-241050--1500-35136075732`
   - Like that of Old Gascon, for `241050--1500`, the '24' key indicates a time period; the '35' key indicates the regional specification as in the tag of Modern Béarnais.

See Fig. 3 for an example of generating a language tag using the web application. When decoding the same language tag, a fragment of the JSON returned is shown in Listing 1. In this code fragment, where each part is shown as `"exists": true`, this means there is a language code for the applicable Part in ISO 639. However, a language code may be in one Part, but not in others. When this is the case, `"exists": false` is returned.

Listing 1: Code Example for a decoded language tag

```
"language": {
    "name": {
        "default": "Occitan (post 1500)"
    },
    "part1": {
        "exists": true,
        "code": "oc",
        "uri": {
            "loc": "http://id.loc.gov/vocabulary/iso639-1/oc",
            "lexvo": "http://lexvo.org/id/iso639-1/oc"
        }
    },
    "londi_iso639_part2": {
        "bibliography": {
            "exists": true,
            "code": "oci",
            "uri": {
                "loc": "http://id.loc.gov/vocabulary/iso639-2/"
            }
        },
        "terminology": {
            "exists": true,
            "code": "oci",
            "uri": {
                "loc": "http://id.loc.gov/vocabulary/iso639-2/oci"
            }
        }
    },
    "part3": {
        "exists": true,
        "code": "oci",
        "uri": {
            "loc": "http://id.loc.gov/vocabulary/iso639-3/oci",
            "lexvo": "http://lexvo.org/id/iso639-3/oci"
        }
    }
},
"privateuse": {
    "language": {
        "key": "02",
        "type": "Identifier from Wikidata",
        "code": "Q35735",
        "uri": "https://www.wikidata.org/wiki/Q35735"
    }
},
```

[14] `https://www.hadw-bw.de/dag.html` [01-12-2019], see also Chambon and Greub (2002, 473).

[15] The date range is supported by Martin Glessgen, director of the DAG [personal communication].

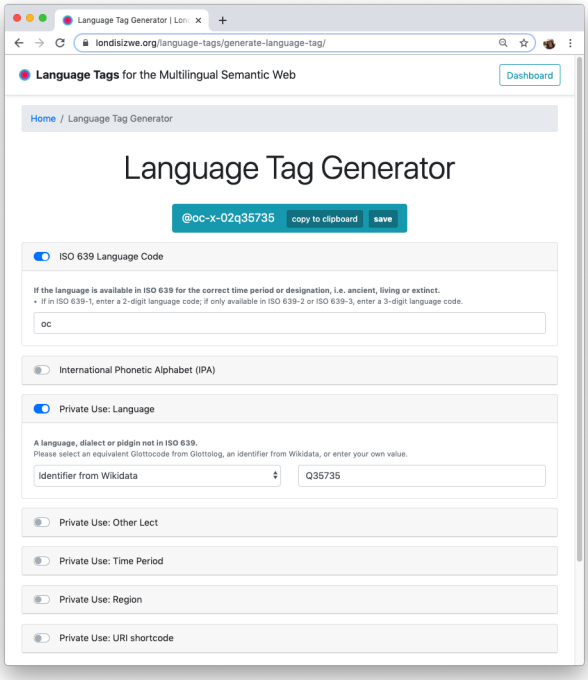

Figure 3: Generating a language tag online

6. Generating URI Shortcodes

The language tags for Gascon and its varieties meets the requirements to identify languages and lects more precisely in time and space. However, the pattern does reveal a compliance problem for the lengthier language tags, in that the sub-tag length exceeds the required length of eight characters as specified by BCP 47. Gillis-Webber and Tittel (2019, 4:11) did acknowledge the issue: BCP 47 itself has *privateuse* portions that are longer than eight characters, see Phillips and Davis (2009, 56; 81) and in BCP 47's appendix `https://tools.ietf.org/html/bcp47#`

appendix-A [29-11-2019]. Although BCP 47 seems to ignore its own specification, to ensure compatibility with resources beyond RDF, limiting the length of the *privateuse* sub-tag to eight characters should ideally be addressed.
To do this, we propose the following:

1. The language tag should be generated as per the pattern in Table 1.
2. In the web application, the user should save this language tag to their profile.
3. Once saved, a URI shortcode will be automatically created for this language tag. The generated shortcode is six characters in length. The user is also able to include an informative annotation of the URI shortcode, with an option to make this annotation publicly visible or private.
4. When using the language tag, use only the URI shortcode, beginning with the `40` key.

To decode the language tag, the user can go to the web application or use the Language Tags API. In both options, the other parts of the encoded language tag from Point 1 will be returned as well. Alternatively, a user can navigate to the following URI: `w3id.org/language-tags/uri/` followed by the URI shortcode. Different format options can also be selected to view: JSON, RDF/Turtle and RDF/NTriples using MoLA.[16]
For **Old Gascon**, the language tag, including the generated URI shortcode, is: `oc-x-02q35735-241050--1500-40000006`.
The language tag can thus be shortened to: `oc-x-40000006`. When the user navigates to the URI: `w3id.org/language-tags/uri/40000006`, the '40' key will be automatically removed and the user will be redirected to the URI: `londisizwe.org/language-tags/uri/000006/page`, where 'page' is the document for the language tag identified by `londisizwe.org/language-tags/uri/000006`. Any user annotations from Point 3, if elected to be publicly visible, will be shown here as well. The same process can be repeated for **Old Béarnais**, rendering the language tag as follows: `oc-x-40000007`.

7. Discussion and Conclusions

The language tags for **Old Gascon** and **Old Béarnais** can then be used, shown in Lines 17 and 19 from Fig. 4, which is an example of Old Gascon lexemes from the entry MOSTRAR of the *Dictionnaire de l'ancien gascon électronique* – DAGél (Glessgen, since 2014).[17]
Although only making use of the URI shortcode in the language tag renders the language tag opaque (when previously it could be considered semi-descriptive), it achieves the aim of enabling languages and lects to be described in a way which supports shared agreement. Furthermore, it is compatible with the existing ISO 639 language code, and is fully compliant with BCP 47. At time of writing, we only offer one URI shortcode service, but as future work, we will attempt to get one or more authoritative organisations to support the language tag, each represented by their own key in the pattern, for eg. '41' and so on.
Our approach to language tagging offers a pattern that is able to accurately represent linguistic variation both through time and space. This pattern supports the efforts to investigate, to provide better access to and, thus, promote linguistic resources, in this instance, of Gascon. Hence, it helps to shed light on the historical background of the socio-cultural being of Southern France's Gascon speaking minority, on linguistic and cultural interactions between the Gascon, Occitan and French speaking populations in the area. However, this is not just limited to Gascon - the same can apply to any lesser-known or endangered language, regional variety or historical language stage.

[16]Model for Language Annotation, `https://ontology.londisizwe.org/mola`, Gillis-Webber et al. (2019).

[17]Please note that the online dictionary entry DAGél MOSTRAR is not yet publicly available due to publication rights (Glessgen and Tittel, 2018).

```
1  @prefix :        <https://dag.adw.uni-heidelberg.de/lemme/> .
2  @prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
3  @refix  rdfs:    <http://www.w3.org/2001/02/rdf-schema#> .
4  @prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
5  @prefix dct:     <http://purl.org/dc/terms/> .
6  @prefix pwn:     <http://wordnet-rdf.princeton.edu/id/> .
7
8  :mostrar a ontolex:LexicalEntry ,
9    ontolex:Word ;
10   rdfs:label "mostrar"@oc-x-40000006 ;
11   lexinfo:PartOfSpeech  lexinfo:Verb ;
12   ontolex:canonicalForm :mostrar_lemma ;
13   ontolex:otherForm     :mostrar_var ;
14   ontolex:evokes :mostrar_lexConcept .
15
16 :mostrar_lemma a ontolex:Form ;
17   ontolex:writtenRep "mostrar"@oc-x-40000006 .
18 :mostrar_var a ontolex:Form ;
19   ontolex:writtenRep "mustrar"@oc-x-40000007 .
20
21 :mostrar_lexConcept a ontolex:LexicalConcept ;
22   ontolex:definition
23   "faire paraître de manière qu'on puisse voir"@fr ;
24   dct:references  pwn:00925764-v .
```

Figure 4: RDF/Turtle code (detail) of DAGél entry *mostrar*.

8. Acknowledgements

The work of Frances Gillis-Webber was financially supported by Hasso Plattner Institute for Digital Engineering.

9. Bibliographical References

Abalain, H. (2007). *Le français et les langues historiques de la France*. Éditions Gisserot, Saint-Brieu.

Baldinger, K. (since 1975). *Dictionnaire onomasiologique de l'ancien gascon – DAG* (fondé par Kurt Baldinger, dirigé par Inge Popelar, puis Nicoline Hörsch / Winkler et Tiana Shabafrouz, sous la direction de Jean-Pierre Chambon, puis Martin Glessgen). De Gruyter [Heidelberger Akademie der Wissenschaften / Kommission für das Altokzitanische und Altgaskognische Wörterbuch], Tübingen / Berlin.

Berners-Lee, T. (2006). *Linked Data*. World Wide Web Consortium. URL: https://www.w3.org/DesignIssues/LinkedData.html [accessed: 01-12-2019].

Chambon, J.-P. and Greub, Y. (2002). Note sur l'âge du (proto)gascon. *Revue de Linguistique Romane*, 66:473–495.

Crystal, D. (2010). *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge.

Field, T. (2009). Présent et passé de la langue de Gascogne. In Guy Latry, editor, *La Voix occitane: Actes du VIIIe Congrès de l'Association Internationale d'Études Occitanes*, pages 2,745–775, Pessac. Presses de l'université de Bordeaux.

Fielding, R. and Reschke, J. (2014). *Internet Engineering Task Force (IETF)*. URL: https://tools.ietf.org/html/rfc7231 [accessed: 28-11-2019].

Gillis-Webber, F. and Tittel, S. (2019). The Shortcomings of Language Tags for Linked Data when Modeling Lesser-Known Languages. In *Proceedings of LDK2019, Leipzig, Germany, 21-22 May 2019, OASIcs, Vol. 70*, pages 4:1–4:15.

Gillis-Webber, F., Tittel, S., and Keet, M. (2019). A Model for Language Annotations on the Web. In Boris Villazón-Terrazas et al., editors, *Knowledge Graphs and Semantic Web. 1st Iberoamerican Conference, KGSWC 2019, Villa Clara, Cuba, June 23-30, 2019, Proceedings*, pages 1–16.

Glessgen, M. and Tittel, S. (2018). Le *Dictionnaire d'ancien gascon électronique* (DAGél). In Roberto Antonelli, et al., editors, *Atti del XXVIII Congresso internazionale di linguistica e filologia romanza (Roma, 18-23 luglio 2016)*, pages 1,805–818, Strasbourg. Société de Linguistique Romane / Éditions de linguistique et de philologie ELiPi, Bibliothèque de Linguistique Romane 15,1.

Glessgen, M.-D., Schmitt, C., Schweickard, W., and Ernst, G. (2008). *Romanische Sprachgeschichte / Histoire linguistique de la Romania (Handbücher zur Sprach- und Kommunikationswissenschaft. Hrsg. von Hugo Steger et al.; Band 23,2)*. De Gruyter, Berlin. URL: http://www.degruyter.com/view/product/36395 [accessed: 01-12-2019].

Glessgen, M. (since 2014). *Dictionnaire de l'ancien gascon – DAGél* (Martin Glessgen en collaboration avec Sabine Tittel). URL: https://dag.adw.uni-heidelberg.de/ [accessed: 01-12-2019].

International Organization for Standardization. (n.d.). *Language codes – ISO 639*. URL: https://www.iso.org/iso-639-language-codes.html [accessed: 29-11-2019].

Manola, F. and Miller, E. (2014). *RDF Primer. W3C Recommendation 10 February 2004*. URL: https://www.w3.org/TR/rdf-primer/ [accessed: 29-11-2019].

Moreux, B. (2004). Béarnais and Gascon today: language behavior and perception. *International journal of the sociology of language (The sociolinguistics of Southern "Occitan" France, revisited)*, 169:25–62.

Moseley, C. (2010). *Atlas of the World's Languages in Danger. Memory of Peoples*. UNESCO Publishing, Paris.

Phillips, A. and Davis, M. (2009). Tags for Identifiying Languages. *BCP*, 47. URL: https://tools.ietf.org/html/bcp47 [accessed: 29-11-2019].

Rey, A., Duval, F., and Siouffi, G. (2007). *Mille ans de langue française. Histoire d'une passion*. Perrin, Paris.

Tittel, S. and Gillis-Webber, F. (2019). Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French. In I. Kosem, et al., editors, *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*, pages 547–569, Brno. Lexical Computing CZ, s.r.o.