Managing Provenance and Versioning for an (Evolving) Dictionary in Linked Data Format

Frances Gillis-Webber[1]
[1]University of Cape Town, Woolsack Drive, Rondebosch, 7701, South Africa
fran@fynbosch.com

Abstract

The *English-Xhosa Dictionary for Nurses* is a unidirectional dictionary with English and isiXhosa as the language pair, published in 1935 and recently converted to Linguistic Linked Data. Using the Ontolex-Lemon model, an ontological framework was created, where the purpose was to present each lexical entry as "historically dynamic" instead of "ontologically static" (Veltman, 2006:6, cited in Rafferty, 2016:5), therefore the provenance information and generation of linked data for an ontological framework with instances constantly evolving was given particular attention. The output is a framework which provides guidelines for similar applications regarding URI patterns, provenance, versioning, and the generation of RDF data.

Keywords: provenance, versioning, multilingualism, lexicography, linked data, ontolex-lemon

1. Introduction

The *English-Xhosa Dictionary for Nurses* (EXDN) is a bilingual dictionary of medical terms, authored by Neil MacVicar, a medical doctor, in collaboration with isiXhosa-speaking nurses. It was the second edition published by Lovedale Press, a South African publisher, in 1935, and as a literary work published in South Africa, it falls under the jurisdiction of the Copyright Act of South Africa, and is now in the public domain, free from any restriction. EXDN is unidirectional; the language pair is English and isiXhosa, with English as the source and isiXhosa the target (Gouws & Prinsloo, 2005; Zgusta, 1971). IsiXhosa (referred here by its endonym) is an indigenous Bantu language from the Nguni language group (S40 in Guthrie's classification) and is an official language of South Africa (Doke, 1954; "Subfamily: Nguni (S.40)", n.d.). Despite it being spoken in South Africa by a large percentage of the population (16.0% counted in the 2011 Census speak it as their L1), it has minority status only (Statistics South Africa, 2012).

Other official South African languages in the Bantu language family (referred hereon as African languages) are: isiNdebele, isiZulu, Sesotho, Sesotho sa Leboa, Setswana, SiSwati, Tshivenḓa, and Xitsonga. In comparison to English, there are limited language resources (LRs) available for these languages, and this, combined with the socio-economic constraints of the speakers, renders these languages under-resourced (Pretorius, 2014; "What is a …", n.d.). Despite English being an ex-colonial language in South Africa, with L1 speakers numbering 9.6%, it is a lingua França with high status, associated with both economic and political power in the country (Ngcobo, 2010; Statistics South Africa, 2012). The African languages listed above, although spoken by the majority, are minority languages and through language shift and death, are at risk of becoming endangered (Pretorius, 2014; Ngcobo, 2010).

In 2009, the first Human Language Technology Audit was conducted (the second audit is currently underway at time of writing (Wilken, personal communication 2017, Dec 12)), with Grover, van Huyssteen and Pretorius identifying the following as issues:

> *"the lack of language resources, limited availability of and access to existing LRs, [and] quality of LRs"*

which hamper the development of new LRs for under-resourced languages (2011, cited in Pretorius, 2014). Although EXDN was published more than seventy-five years ago, as a LR for an under-resourced language, its content is still valuable. Linked Data is a simple data model with an interoperable format and by publishing lexicographic resources, particularly lesser-known resources such as EXDN, in Linked Data, it enables the "aggregation and integration of linguistic resources", which can serve as an aid for the future development of new and existing LRs (Gracia, 2017).

Using the Ontolex-Lemon model, an ontological framework was created, where the purpose was to present each lexical entry as "historically dynamic" instead of "ontologically static" (Veltman, 2006:6, cited in Rafferty, 2016:5), therefore the provenance information and generation of Linked Data for an ontological framework with instances constantly evolving was given particular attention.

The rest of the paper is organised as follows: in Section 2, the structure of the dictionary is briefly described; in Section 3, the URI strategy is discussed; in Sections 4 and 5, the description of resources, provenance for lexical entries and the lexicons are considered, and the versioning and generation of Linked Data is presented. The conclusions of the paper are presented in Section 6.

2. The Structure of the Dictionary

The frame structure of a dictionary is typically composed of the central list, with front and back matter texts (Gouws & Prinsloo, 2005); however, the frame structure of EXDN consists of a central list, with front matter texts only. The central list of EXDN is represented by the Roman alphabet, with each letter acting as a guiding element for a series of article stretches (Gouws & Prinsloo, 2005).

EXDN can also be described according to its macrostructure and microstructure. EXDN's macrostructure comprises a lemmatised list in the source language only: English - ordered alphabetically with a

singular and plural lemmatisation of nouns (Gouws & Prinsloo, 2005). A dictionary's microstructure pertains to the structure of each article (lexical entry), with the lemma serving as a guiding element for each (Gouws & Prinsloo, 2005). In the case of EXDN, each article comprises one of the following (or a combination thereof): lexicographic definition, a translation, or a cross-reference entry. If the article has a single target language item, shown by a single word, then it is presumed that the article is a translation equivalent, with full equivalence (Gouws & Prinsloo, 2005). However, if the article has a lexicographic definition in the target language, then zero equivalence is presumed (Gouws & Prinsloo, 2005).

3. The URI Strategy

Archer, Goedertier and Loutas have defined a URI as "a compact sequence of characters that identifies an abstract or physical resource" and it "can be further classified as a locator, a name, or both" (2012).

A key set of principles have been identified for URIs:

- URIs should be:
  - short,
  - stable,
  - persistent, and
  - human-friendly (Archer, Goedertier & Loutas, 2012; Hogan et al., 2012; Wood et al., 2014).
- URIs should be HTTP(S) URIs (Berners-Lee, 2006; Hogan et al., 2012).
- The identifier portion of a URI should be:
  - unique,
  - and unambiguous (Simons & Richardson, 2013; Keller et al., 2011).
- URIs should be dereferenceable, with a representation returned when a human or software agent navigates to the URI (Heath & Bizer, 2011; Hyvönen, 2012).
- URIs should differentiate between the resource, and the document which describes a resource (Van Hooland & Verborgh, 2014; Heath & Bizer, 2011).

In the sub-sections that follow, fragment identifiers, URI patterns, and resource identifiers are discussed in more detail.

3.1 Fragment Identifiers

Fragment identifiers are an optional part of the URI, positioned at the end, and are of the pattern "#example". Although the usage of fragment identifiers have been cautioned against by Wood et al., primarily because web servers do not process the fragment, they are widely used in vocabularies, where "the vocabulary is often served as a document and the fragment is used to address a particular term within that document" (2014). Within the context of identifying sub-resources in relation to the parent resource, fragment identifiers can be useful, as they can clearly show a hierarchical relationship with the parent resource (however, deeper levels cannot be indicated).

According to Sachs and Finin, the URI should resolve "not to the address, but to all known information about the resource" (2010); from this one can infer that when information for a sub-resource is returned, then information for the parent resource should also be returned. Conversely, when information for a parent resource is returned, information of any sub-resources should also be returned. By doing this, the need to have a separate document to describe the parent resource and each of the sub-resources is not necessary, as one document can be used to describe the parent resource and any sub-resources.

Additionally, when publishing Linked Data and versioning is employed, by using fragment identifiers to identify sub-resources within the same document, redundancy can be reduced.

3.2 The URI Pattern

When working with EXDN data, the following use cases were determined (Gillis-Webber, 2018):

U1: *A URI which identifies a resource*

U2: *A URI which identifies a sub-resource in relation to the parent resource*

U3: *A URI which identifies a version of the resource*

U4: *A URI which identifies a version combined with a sub-resource*

U5: *A URI which identifies a document describing the resource in **U1***

U6: *A URI which identifies a document describing the resource in **U3***

A pattern for a URI has been recommended by Archer et al. (2012):

```
http://{domain}/{type}/{concept}/{reference}
```

Where:

- {domain} is the host,
- {type} is the resource (for eg. *id*) being identified,
- {concept} refers to a real world object or a collection, and
- {reference} is the local reference for the resource being identified.

When using the *lemon* model (a previous iteration of the Ontolex-Lemon model), Gracia and Vila-Suero developed a set of guidelines for publishing Linked Data for bilingual dictionaries, and they too proposed the same pattern as Archer et al. (2015). As an example, for the lexical entry "bench", the URI is as follows:

E1: `http://linguistic.linkeddata.es/id/apertium/lexiconEN/bench-n-en`

Where:

- *linguistic.linkeddata.es* is the host,
- *id* is the resource,
- *apertium* is the collection,
- *lexiconEN* is the source lexicon,
- *bench-n-en* is the reference.

When considered from a user perspective, the human-friendliness of **E1** can be evaluated accordingly:

- *id* is not particularly informative and could be deemed redundant;
- although specifying the collection (*apertium*) is useful, should a dataset from another collection be merged with the existing dataset, if there are shared lexical entries between both collections, this will result in URIs which are incongruently defined;
- both the lexicon and the reference are identifiable as English, thus *lexiconEN* could also be deemed redundant.

Ontolex-Lemon (and *lemon* as well) requires the lexical entries in a lexicon to be the same language. If modelling two languages, then the lexical entries of each language would be contained within their own lexicon, with translation relations explicitly defined between the corresponding lexical entries or their senses, using the *vartrans* module ("Final model specification", n.d.). BabelNet was also modelled on *lemon*, and by 2015 it had 271 lexicons, one for each of the languages it supported; Flavi et al. remarked on this saying *lemon* requires "us to work on a language-by-language basis, whereas in BabelNet this distinction does not need to be made explicit".

Continuing with the example lexical entry "bench", in BabelNet, the URI is as follows:

E2: `http://babelnet.org/rdf/bench_n_EN`

There should be a separation between the URIs and the model used to describe the lexical data. If the model should change, the persistence and longevity of the URIs should not be impacted, and as a result, a "URI should be agnostic of the selected model" (Gillis-Webber, 2018). For **E1** and **E2**, both the references (*bench-n-en* and *bench_n_EN* respectively) have been encoded with additional information by appending the lemma with the language shortcode and an abbreviated form of part-of-speech (POS), and by doing this, the URIs for the two examples are identifiable to be of the English language with POS noun.

E1 could therefore be revised to:

`http://linguistic.linkeddata.es/entry/bench-n-en`

And for a lexicon:

`http://linguistic.linkeddata.es/lexicon/en`

For each of the six use cases identified for EXDN at the beginning of Section 3.2, the application of this simplified pattern has continued, and below, the pattern of each use case is provided, followed by a short description thereof, as well as an associated example from Londisizwe.org, the multilingual online dictionary derived from the EXDN dataset.

A URI which identifies a resource has the form (Gillis-Webber, 2018):

U1: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}*

Where:

- {http(s):} is the http: or https: scheme
- {Base URI} is the host
- {Resource Path}, for example, *entry* for a lexical entry, and *lexicon* for a lexicon
- {Resource ID}, for example, *en-n-abdomen*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen`

A URI which identifies a sub-resource in relation to the parent resource has the form (Gillis-Webber, 2018):

U2: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}#{Fragment ID}*

Where:

- {Fragment ID} is the fragment identifier, for example, *sense1*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen#sense1`

The resource identifier, described in **U1**, will be unique relative to the resource path. The fragment identifier will be unique relative to the resource identifier.

A URI which identifies a version of the resource has the form (Gillis-Webber, 2018):

U3: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}/{Version ID}*

Where:

- {Version ID} is the version identifier, for example, *2017-09-19*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen/2017-09-19`

As the sub-resource is identified in relation to the parent resource, any change to the sub-resource would result in a change to the URI of the parent resource.

Therefore, a URI identifying a sub-resource when employing the use of versioning has the form (Gillis-Webber, 2018):

U4: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}/{Version ID}#{Fragment ID}*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen/2017-09-19#sense1`

For a resource, each version should be dereferenceable, and should remain so even as newer versions of the same resource are published. Like that of the fragment identifier, the version identifier is unique to the resource

identifier. The use case **U1** will resolve to the latest version available for that resource (Archer et al., 2012).

A URI which identifies a document describing the resource in **U1** has the form (Gillis-Webber, 2018):

U5: *{http(s):}//{Base URI}/{Document}/{Resource Path}/{Resource ID}*

Where:

- Using content negotiation, {Document} refers to the HTML page, for example, *page*, or to the RDF representation, for example, *rdf*, using any form of serialisation.

Corresponding examples are:

`https://londisizwe.org/page/entry/en-n-abdomen`

`https://londisizwe.org/rdf/entry/en-n-abdomen`

A URI which identifies a document describing the resource in **U3** has the form (Gillis-Webber, 2018):

U6: *{http(s):}//{Base URI}/{Document}/{Resource Path}/{Resource ID}/{Version ID}*

An example URI is:

`https://londisizwe.org/rdf/entry/en-n-abdomen/2017-09-19`

In the context of EXDN, a document which describes **U2** (or **U4**) is not necessary, and instead it resolves to **U5** (or **U6**).

3.3 Resource Identifiers

The human-friendliness of URIs has been suggested in the literature, with frequent references thereto: such as URIs should be "user-friendly" (Archer, Goedertier & Loutas, 2012), "human readable" (Hogan et al., 2012), "meaningful" (Villazón-Terrazas et al., 2012), and "natural keys" should be used (Wood et al., 2014; Heath & Bizer, 2011). Defined by Labra Gayo, Kontokostas and Auer (n.d.) as "descriptive URIs", and as "meaningful URIs" by Vila-Suero et al. (2014), this type of URI is generally used "with terms in English or in other Latin-based languages" (Labra Gayo, Kontokostas & Auer, n.d.).

Labra Gayo et al. defines "opaque URIs" as "resource identifiers which are not intended to represent terms in a natural language", with it suggested by both Labra Gayo, Kontokostas and Auer (n.d.), and Vila-Suero et al. (2014) that in a multilingual context, using opaque URIs is preferable so as to avoid language bias. By doing so within the context of the Semantic Web, Vila-Suero et al. argue that this is acceptable, as "resource identifiers are intended for machine consumption so that there is no need for them to be human readable" (2014).

Within the larger context of the Semantic Web, this view may be accurate as data models are mostly language-agnostic (Ehrmann, 2014), however in the context of Linked Data, it is in opposition to a fundamental principle thereof: a URI should be dereferenceable, to be looked up by either a web browser for human consumption or a software agent (Hyvönen, 2012).

Due to the localisation of this study within South Africa and its languages being Latin-based, a pragmatic approach was taken with regards to the URIs: descriptive URIs were used, using English, however in a similar approach to Babelnet, opaque URIs were used when modelling the lexical concepts (Flati et al., 2015).

For lexical entries, a similar approach as that used in **E1** was taken for the resource identifiers, however the elements were reordered to aid programmatic extraction (should it be required):

`{Language Code}-{POS}-{Lemma}`

Where:

- {Language Code} is the lowercase form of the language shortcode, using ISO 639-1, and if none available, then ISO 639-2 (or ISO 639-3) will be used
- {POS} is an abbreviated form of POS, described in English
- {Lemma} is the lowercase form of the lemma, with underscores replacing any hyphens or spaces and any diacritics are removed

For a lexical entry, a constraint of the Ontolex-Lemon model is that it can be associated with exactly one POS and exactly one language ("Final model specification", n.d.). For lexical entries which may share the same lemma, such as:

isiXhosa: *isibindi*
isiZulu: *isibindi*

to avoid potential collision, it was considered best for the EXDN dataset to include the language shortcode and the abbreviated POS in the identifier as well, thus allowing for the easy extensibility of the existing dataset to additional languages. Thus for the two lexical entries above, their identifiers would be as follows:

isiXhosa: `xh-n-isibindi`
isiZulu: `zu-n-isibindi`

For a lexicon, the resource identifier takes the form (shown here including the resource path):

`{Resource Path}/{Language Code}`

In combination with the resource path, the resource identifier should adequately identify the lexical entry (or lexicon), thus allowing for any language to be represented (with the exception of the written form of sign languages, which can conceivably be any language) (Gillis-Webber, 2018).

4. The Description of Resources

As previously mentioned, when returning information for a resource and any of its sub-resources, the information returned should not be limited to describing these resources, the inclusion of the following additional information could be considered as well (Gillis-Webber, 2018):

- *A description of related resources;*

- *A description of the metadata of the resource (for example, provenance and version);*
- *A description of the dataset which contains the resource (Heath & Bizer, 2011:45).*

In the case of EXDN, when publishing the information for a lexicon which resolves, for example, to the URI `https://londisizwe.org/lexicon/en`, it was not considered practical to include information of the related resources, particularly for each lexical entry. However, when publishing the information for a lexical entry which resolves, for example, to the URI `https://londisizwe.org/entry/en-n-abdomen`, it was considered necessary, and the following additional information is thus included (Gillis-Webber, 2018):

- *Description of the document which describes the lexical entry,*
- *Metadata of the lexical entry,*
- *Provenance information of the lexical entry,*
- *Identification of the lexicon to which the entry belongs,*
- *Brief description of other lexical entries, resources and ontology entities related to the lexical entry.*

5. Modelling Provenance & Versioning

According to Di Maio (2015), knowledge is "partial/incomplete/imperfect, with very few exceptions". Linked Data is about relationships, and when considered within the context of Linguistics, datasets of different lexicons can be interlinked, thus allowing for the extension of an existing lexicon; for under-resourced languages, this can be a powerful notion (Berners-Lee, 2009; McCrae et al., 2012). According to Bouda and Cysouw (2012), when retrodigitising language resources, the encoding thereof is not the challenge, but rather "the continuing update, refinement, and interpretation" of the dataset, and with each change, providing for traceability. Like RDF datasets, ontologies and vocabularies are not static, and they too evolve over time (Hyvönen, 2012). This change can be attributed to factors such as error correction, the addition of concepts and properties to the underlying model, as well as change out in the world, and our understanding thereof (Hyvönen, 2012).

As mentioned in Section 2, within the context of EXDN, until established otherwise, then full equivalence is presumed if the article has a single target language item, and if anything more than a single target language, then it is presumed the article is a lexicographic definition and there is zero equivalence (Gouws & Prinsloo, 2005).

Google's Cloud Translation API[1] was used to translate the isiXhosa texts, with English selected as the target language. There are two models available: Phrase-Based Machine Translation model (PBMT) and Neural Machine Translation model (NMT), and using each model, an article was translated ("Translating text", n.d.). As an example, the article *stomach*, which has the isiXhosa text of "Uluusu lomntu.", when translated on 2017-09-17 20:00:31 GMT+2, yielded the following:

PBMT: A person's skin.
NMT: Homosexuality.

[1] https://cloud.google.com/translate/

There are several possibilities for this: (1) the source data contains errors, (2) the source data is so outdated that it is not possible to translate this accurately, or (3) there are not enough existing language pairs within the Cloud Translation API to accurately translate the text ("Cloud translation API", n.d.). According to Google's website, the Cloud Translation API undergoes continuous updates ("Cloud translation API", n.d.) so although it is intended to periodically repeat the translation process for the EXDN dataset, for now, the translated texts are not used for disambiguation purposes.

Continuing with the article *stomach*, when the lexical entry with the identifier `en-n-stomach` was first published in 2017, its only sense (`en-n-stomach#sense1`) was linked to a lexical concept (`https://londisizwe.org/concept/000000007`) which had a language-tagged lexicographic definition `"Uluusu lomntu."@xh`, and it was set as a concept of the DBpedia resource: `http://dbpedia.org/resource/Stomach`. However, after consultation with isiXhosa mother tongue speakers in early 2018, the following was determined:

- "uluusu" was incorrectly spelt in EXDN (it should have been "ulusu"),
- the equivalent of *stomach* is also *isisu*,
- the meaning (gloss) of "ulusu lomntu" is "a person's stomach", however it was difficult to determine if the text should remain a lexicographic definition or if it should become a lexical entry with "ulusu lomntu" as the lemma.

As a result of this new information, the following changes were implemented:

- For the lexical entry `en-n-stomach`, the spelling mistake was corrected in the lexical concept.
- The lexical entry `xh-n-isisu` already existed, however another sense was added (`xh-n-isisu#sense2`), and it was linked to the same lexical concept.

Because there is a shared conceptualisation between `https://londisizwe.org/entry/en-n-stomach#sense1` and `https://londisizwe.org/entry/xh-n-isisu#sense2`, they are deemed to be equivalent.

As the purpose of digitising EXDN and converting its dataset to Linked Data is to enable its reuse by external resources, it is important that any changes are accurately recorded, by way of versioning, with provenance information included as well. The lexical entry `xh-n-isisu` had a change to one of its senses (a sub-resource), and the lexical concept `000000007` changed as well, so there is now a new version for each. As there were not any insertions or deletions for the English and isiXhosa lexicons, these remained unchanged. In the event the lexical entries had to be reviewed again, it is expected they would be subject to further refinement.

As an aside, the Cloud Translation API was used again (2018-03-02 20:36:48 GMT+2), this time with the corrected text "Ulusu lomntu.". PBMT remained unchanged, however NMT returned the following translation: "Human skin.". It was also repeated for the original source text, and those translations remained unchanged from 2017-09-17.

5.1 Versioning

Versioning is used by Babelnet, although it is applied globally for their BabelNet-lemon schema description, with Flati et al. acknowledging that "maybe a more sophisticated infrastructure would be needed in order to express more complex versioning description needs" (2015). When the generation and publication of RDF data for the Apertium Bilingual Dictionaries was detailed by Gracia et al., versioning was not included in the discussion (n.d.). Although briefly mentioned by McCrae et al. (2012), Gracia et al. (n.d.), Eckart et al. (2012), van Erp (2012), and De Rooij et al. (2016), it does not appear that versioning has been discussed further within the domain of Linguistic Linked Data, and in the context of vocabularies used by Babelnet, Flati et al. commented that changes are unaccounted for "and this aspect might thus be investigated in more detail in the [near] future by the whole community" (2015).

When describing the generation of RDF for the Apertium Bilingual Dictionaries, Gracia et al. talked of three RDF files: one per lexicon, and the third for the translations (n.d.). From this, the author inferred that if versioning was implemented, it would be done at file-level, in a similar approach to that taken by BabelNet. However, in the context of EXDN, it was felt that publishing only at the lexicon-level could become unmanageable over time, particularly on a 24-hour publishing schedule, and instead it would be more practical to implement versioning at the lexical entry-level as well. Versioning at the lexicon-level is also done, but a file only includes the changes from the previously published version, and any additional information of the lexical entries, beyond the resource identifier, is excluded. For each version of a lexical entry, the file contains: all information of the lexical entry, its senses, and translation relations for which any of its senses is the source.

Thus, the following components have been identified for the versioning of EXDN (Gillis-Webber, 2018):

- *Versioned URIs for lexicons, lexical entries, and senses*
- *Provenance metadata to describe the versions, with the latest version mapping to previous versions (Van Erp, 2012), and*
- *The generation of files, one for each version of the lexical entries and lexicons.*

Within the context of EXDN, lexical concepts are modelled as a shared conceptualisation between senses, and they can be thought of as similar to that of a WordNet *synset*, however, where WordNet models sets of similar terms, lexical concepts model sets of equivalent senses across languages (Bosque-Gil et al, 2015). Although the lexical concepts are hosted on the same domain, they are stored within a sense inventory called *Londisizwe Concepts for Senses*[2] – this is considered to be a standalone inventory, and as a result, it is not described further here, although the same principles for versioning do apply (Gillis-Webber, 2018).

Section 3.2 introduced versioned URIs, with the use cases: **U3** and **U4**. Modelling provenance, and the generation and publication of Linked Data are discussed in the sections that follow.

[2] https://londisizwe.org/concept

5.2 Modelling Provenance for a Lexical Entry, its Senses, and Translation Relations

The W3C Provenance Working Group defines provenance ("PROV-O", 2013):

as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing.

A factor contributing to the reuse of a RDF dataset, either by linking or by using the downloaded data, is trust – trust in the repository supplying the data, and trust in the data itself (Faniel & Yakel, 2017). By documenting the provenance of data using a systematic schema, provenance provides a trust marker (essential in an open environment like the web); and within the context of EXDN, provenance information is documented using the PROV Ontology, DCMI Metadata terms, and versioned URIs (Faniel & Yakel, 2017; "PROV-O", 2013; Tennis, 2007; Flati et al., 2015).

The metadata used to describe the EXDN dataset is as follows:

- Each lexical entry, sense, and translation relation is identified as a `prov:Entity`.
- The `prov:generatedAtTime` property is recorded for each.
- The date a lexical entry, sense or translation relation is changed is recorded using `dct:modified`.
- The person or organisation responsible for creating the lexical entry or sense is identified using `dct:creator`.
- The source from which a lexical entry is primarily derived is identified using the `prov:hadPrimarySource` property.
- The other sources from which a lexical entry, sense or translation relation is derived, is identified using the `dc:source` property.
- One or more contributors (a person, an organisation or a service) for a lexical entry, sense or translation relation is identified using `dct:contributor`.
- The licensing agreement for a lexical entry is identified using `dct:license`, and Creative Commons is used for the licensing.
- For a lexical entry, `dct:isPartOf` is used to denote inclusion of a lexical entry in a lexicon, and inclusion of a sense in a lexical entry.
- For a translation relation, `dct:hasPart` is used to identify both the source and target language.
- For a lexical entry, `owl:sameAs` is used to indicate that **U1** is the same as the latest version of **U3**.
- For a sense or translation relation, `owl:sameAs` is used to indicate that **U2** is the same as the latest version of **U4**.
- For a lexical entry, sense or translation relation, the version is indicated using `owl:versionInfo`.
- For a lexical entry, sense or translation relation, `dct:hasVersion` is used to show the previously generated versions, using the versioned URIs (**U3** for lexical entries and **U4** for senses and translation relations).

The generated RDF for version two of the lexical entry `xh-n-isisu` follows below. The lexical concept for `000000001` is also shown for reference purposes.

```
@prefix :      <https://londisizwe.org/> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dbr:   <http://dbpedia.org/resource/> .
@prefix dct:   <http://purl.org/dc/terms/> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix lcnaf: <http://id.loc.gov/authorities/names/> .
@prefix lcsh:  <http://id.loc.gov/authorities/subjects/> .
@prefix lexinfo:
              <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@prefix mesh:  <http://id.nlm.nih.gov/mesh/> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix prov:  <http://www.w3.org/ns/prov#> .
@prefix pwn:   <http://wordnet-rdf.princeton.edu/rdf/id/> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix void:  <http://rdfs.org/ns/void#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

<https://londisizwe.org/rdf/entry/xh-n-isisu>
    rdfs:label   "RDF document for the lexical entry:
                 isisu, n (isiXhosa)"@en ;
    rdf:type     foaf:Document ;
    foaf:primaryTopic :entry/xh-n-isisu .

:entry/xh-n-isisu
    a ontolex:LexicalEntry , ontolex:Word , prov:Entity ;
    lexinfo:partOfSpeech  lexinfo:Noun ;
    dct:language <http://id.loc.gov/vocabulary/iso639-2/xho> ,
                 <http://lexvo.org/id/iso639-1/xh> ;
    dct:identifier :entry/xh-n-isisu ;
    rdfs:label    "isisu"@xh ;
    ontolex:canonicalForm :entry/xh-n-isisu#lemma ;
    ontolex:sense :entry/xh-n-isisu#sense1 ,
                  :entry/xh-n-isisu#sense2 ;
    dct:subject   mesh:D000005 ;
    ontolex:denotes dbr:Abdomen , dbr:Stomach ;
    ontolex:evokes :concept/000000001 ;
    dct:isPartOf  :lexicon/xh ;
    dct:license
    <http://creativecommons.org/publicdomain/mark/1.0/> ;
    prov:hadPrimarySource "The English-Xhosa Dictionary for
                  Nurses"@en ;
    dct:creator   <https://londisizwe.org> ;
    prov:generatedAtTime "2018-01-
                  10T05:00:00Z|+02:00"^^xsd:dateTime ;
    dct:modified  "2018-01-10"^^xsd:date ;
    owl:versionInfo "2018-01-10"^^xsd:string ;
    owl:sameAs     :entry/xh-n-isisu/2018-01-10 ;
    owl:hasVersion :entry/xh-n-isisu/2017-09-19 ,
                   :entry/xh-n-isisu/2018-01-10 .

:entry/xh-n-isisu#lemma
    a             ontolex:Form ;
    ontolex:writtenRep  "isisu"@xh .

:entry/xh-n-isisu#sense1
    a             ontolex:LexicalSense , prov:Entity ;
    ontolex:isLexicalizedSenseOf :concept/000000001 ;
    dct:identifier :entry/xh-n-isisu#sense1 ;
    dct:isPartOf  :entry/xh-n-isisu ;
    dct:creator   <https://londisizwe.org> ;
    prov:generatedAtTime "2018-01-
                  10T05:00:00Z|+02:00"^^xsd:dateTime ;
    dct:modified  "2018-01-10"^^xsd:date ;
    owl:versionInfo "2018-01-10"^^xsd:string ;
    owl:sameAs    :entry/xh-n-isisu/2018-01-10#sense1 ;
    owl:hasVersion  :entry/xh-n-isisu/2017-09-19#sense1 ,
                    :entry/xh-n-isisu/2018-01-10#sense1 .

:entry/xh-n-isisu#sense2
    a             ontolex:LexicalSense , prov:Entity ;
    ontolex:isLexicalizedSenseOf :concept/000000007 ;
    dct:identifier :entry/xh-n-isisu#sense2 ;
    dct:isPartOf  :entry/xh-n-isisu ;
    dct:creator   <https://londisizwe.org> ;
    prov:generatedAtTime "2018-01-
                  10T05:00:00Z|+02:00"^^xsd:dateTime ;
    owl:versionInfo "2018-01-10"^^xsd:string ;
    owl:sameAs    :entry/xh-n-isisu/2018-01-10#sense2 ;
    owl:hasVersion  :entry/xh-n-isisu/2018-01-10#sense2 .

:concept/000000001
    a              skos:Concept , ontolex:LexicalConcept ;
    ontolex:lexicalizedSense :entry/en-n-abdomen#sense1 ;
    ontolex:lexicalizedSense :entry/xh-n-isisu#sense1 ;
    owl:sameAs     pwn:05564576-n ;
    owl:sameAs     mesh:M000005 ;
    dct:subject    mesh:D000005 ;
    ontolex:isConceptOf  dbr:Abdomen .
```

Figure 1: Modelling of a lexical entry

5.3 Modelling Provenance for a Lexicon

Using the same principles from the previous sections, as well as the *lime* module from Ontolex-Lemon, the metadata of EXDN's isiXhosa lexicon is described below in RDF. The metadata only serves to describe the lexicon, and when a lexical entry is inserted or removed from a lexicon is not described. However, PROV-Dictionary[3], published by the W3C Provenance Working Group in 2013 as an extension to PROV, "introduces a specific type of collection, consisting of key-entity pairs", thus allowing for the change of lexical entries in a lexicon, as members of a collection, to be expressed as well ("PROV-Dictionary: Modeling provenance …", 2013).

The generated RDF for version three of the lexicon `xh` follows below:

```
@prefix :      <https://londisizwe.org/> .
@prefix lime:  <http://www.w3.org/ns/lemon/lime#> .
@prefix dct:   <http://purl.org/dc/terms/> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix prov:  <http://www.w3.org/ns/prov#> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix void:  <http://rdfs.org/ns/void#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

<https://londisizwe.org/rdf/lexicon/xh>
    rdfs:label   "RDF document for the lexicon: isiXhosa"@en ;
    rdf:type     foaf:Document ;
    foaf:primaryTopic :lexicon/xh .

:lexicon/xh
    a lime:Lexicon , void:Dataset ,
      prov:Dictionary , prov:Collection , prov:Entity ;
    lime:language  "xh" ;
    dct:language <http://id.loc.gov/vocabulary/iso639-2/xho> ,
                 <http://lexvo.org/id/iso639-1/xh> ;
    dct:identifier :lexicon/xh ;
    lime:lexicalEntries "1"^^xsd:integer ;
    lime:linguisticCatalog
          <http://www.lexinfo.net/ontologies/2.0/lexinfo> ;
    dct:description "Londisizwe.org - isiXhosa lexicon"@en ;
    dct:creator    <https://londisizwe.org> ;
    prov:generatedAtTime "2018-01-
                  15T06:00:00Z|+02:00"^^xsd:dateTime ;
    dct:modified   "2018-01-15"^^xsd:date ;
    owl:versionInfo "2018-01-15"^^xsd:string ;
    owl:sameAs     :lexicon/xh/2018-01-15 ;
    owl:hasVersion :lexicon/xh/2017-09-19 ,
                   :lexicon/xh/2018-01-12 ,
                   :lexicon/xh/2018-01-15 ;
    dct:references :lexicon/en ;
    void:dataDump  <https://londisizwe.org/data/xh-
                   lexicon/2018-01-15> .

:lexicon/xh/2018-01-12
    a              prov:Dictionary .

:lexicon/xh/2018-01-15
    a              prov:Dictionary ;
    prov:derivedByRemovalFrom :lexicon/xh/2018-01-12 ;
    prov:qualifiedRemoval [
        a          prov:Removal ;
        prov:dictionary :lexicon/xh/2018-01-12 ;
        prov:removedKey "xh-n-ulusu_lomntu"^^xsd:string ;
    ] ;
.
```

Figure 2: Modelling of a lexicon

Where:

- Lines 36 – 37: the previous version is identified as a dictionary. There were two dictionary entries, although those entries are not listed here, instead they would have been listed in the file of the previously published URI: `https://londisizwe.org/lexicon/xh/2018-01-12`

[3] https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/

- Lines 39 – 40: the current version is identified as a dictionary.
- Line 41: states that the current version was derived from the previous version.
- Lines 42 – 46: indicates the key that was removed. There is now only one lexical entry, `xh-n-isisu`, in the isiXhosa lexicon.

The class `prov:Dictionary` is defined as "an entity that provides a structure to some constituents, which are themselves entities. These constituents are said to be members of the dictionary", and the concept of 'dictionary' can be extended to include "a wide variety of concrete data structures, such as maps or associative arrays" ("PROV-Dictionary: Modeling provenance …", 2013). Within the context of EXDN, while `prov:Dictionary` has only been applied to lexicons, it could conceivably also be applied to lexical entries and lexical concepts – both of which are containers, with each having senses as its members. While this has not yet been explored for the EXDN dataset, it is work that will be considered in the future.

5.4 Generation and Publication of Linked Data

In a similar vein to versioning, the generation and publication of RDF data is only briefly mentioned in the literature (Vila-Suero et al., 2014; Ehrmann et al., 2014; Gracia et al., n.d.), although for BabelNet, Ehrmann et al. did talk of RDF dump files (which no longer seem to be available for download). For the RDF files discussed in Section 5.1 for the Apertium Bilingual Dictionaries, Gracia et al. (n.d.) talked of loading them into a Virtuoso[4] triple store, with a SPARQL endpoint to access the RDF data, as well as the development of a Linked Data interface using Pubby[5]. The topic was explored further by Gracia in a presentation in 2017, recommending the use of a SPARQL store, with "a mechanism to make [our] URIs dereferenceable: through a common web server (as files)", or by making use of a Linked Data interface. According to Heath and Bizer (2011), storing static RDF files on a web server is "the simplest way to publish Linked Data", and within the context of EXDN, this was the selected route. A Dictionary Writing System was custom-developed for the purpose of maintaining the EXDN dataset, with automated processes implemented for file generation.

Because of the versioning requirements listed in the previous section, the following approach to publication is taken:

- When a lexical entry (or its senses or translation relations of which one of the senses of the lexical entry is the source) changes, a new file in the various formats required is generated. **U1** always point to the latest version of the lexical entry. This is an automated task, scheduled to run daily at 5AM.
- Lexical entries are members of the lexicon collection, and if there are any changes to the members (insertions or deletions), then a new version of the lexicon file is generated, using the same principle as that described for lexical entries. This process is repeated per lexicon.
- The files representing the latest version of the lexicon and its lexical entries are manually merged and compressed to create a data dump. It is planned to automate this process in the future. A SPARQL endpoint is currently not available, although it is planned to trial Dydra[6], a cloud-hosted RDF platform ("Dydra", 2011).

6. Conclusion

Although EXDN was published in 1935, once the dataset is fully converted to Linked Data, it will continue to evolve: with the identification of additional resources to link to; by merging with other LRs; as well as the planned implementation of a crowdsourcing approach to correct, change, and add lexicographic definitions, cross-reference entries, translations, senses, and annotations to lexical entries in multiple African languages. Within the context of EXDN, provenance and versioning has thus been identified as essential components whilst converting the dictionary to Linguistic Linked Data, as well as for its on-going improvements thereafter.

Furthermore, the lemmatisation approach for African languages, as well as annotations within a multilingual environment were modelling challenges identified by the author whilst working with the EXDN dataset. Likewise, the representation of hierarchy in RDF, be it in the form of sub-senses, or inflection, with multiple affixes attached to a word stem, has been identified as a modelling challenge by Gracia, Kernerman and Bosque-Gil (2017). Both a lexicography module and a morphology module for the Ontolex-Lemon model is in progress with the Ontology-Lexica Community Group, and when implemented, it is expected that the modelling of EXDN's lexical entries and senses may change (Bosque-Gil, 2017; McCrae & Gracia, 2017). Although the Ontolex-Lemon model takes a modular approach, as its range extends, provenance and versioning will be of importance so that any change to the RDF representation of data is accurately recorded.

7. Acknowledgements

Thank you to the reviewers for the kind advice and suggestions on improving this paper.

8. Bibliographical References

Archer, P., Goedertier, S. & Loutas, N. (2012). *D7.1.3 – Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC.* Available: https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf [2017, December 26].

Berners-Lee, T. (2006). *Linked data*. Available: https://www.w3.org/DesignIssues/LinkedData.html [2017, December 25].

Berners-Lee, T. (2009). *Tim Berners-Lee: the next web* [Video file]. Available: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html [2017, April 15].

[4] https://virtuoso.openlinksw.com/

[5] http://wifo5-03.informatik.uni-mannheim.de/pubby/

[6] https://dydra.com

Bosque-Gil, J. (2017). Linked data and dictionaries [Seminar]. 2nd Summer Datathon on Linguistic Linked Open Data. 27 June.

Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G. & Montiel-Ponsoda, E. (2015). Applying the OntoLex model to a multilingual terminological resource. In *The semantic web: ESWC 2015 satellite events*. F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker & A. Zimmerman, Eds. 283-294.

Bouda, P. & Cysouw, M. (2012). Treating dictionaries as a linked-data corpus. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 15-24.

Cloud translation API: dynamically translate between thousands of available language pairs. n.d. Available: https://cloud.google.com/translate/ [2018, February 28].

De Rooij, S., Beek, W., Bloem, P., van Harmelen, F. & Schlobach, S. (2016). Are names meaningful? Quantifying social meaning on the semantic web. In *The Semantic Web: ISWC 2016.* P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck & Y. Gil, Eds. 184-199.

Di Maio, P. (2015). Linked data beyond libraries. In *Linked data and user interaction.* H.F. Cervone, L.G. Svensson, Eds. Berlin: Walter de Gruyter GmbH. 3-18.

Doke, C.M. 1954. *The Southern Bantu languages.* London: International African Institute.

Dydra. (2011). Available: https://www.w3.org/2001/sw/wiki/Dydra [2018, January 3].

Eckart, K., Riester, A. & Schweitzer, K. (2012). A discourse information radio news database for linguistic analysis. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 65-76.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P. & Navigli, R. (2014). *Representing multilingual data as linked data: the case of Babelnet 2.0.* Available: http://wwwusers.di.uniroma1.it/~navigli/pubs/LREC_2014_Ehrmannetal.pdf [2017, December 27].

Faniel, I.M. & Yakel, E. (2017). Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. In *Curating research data: Practical strategies for your digital repository*. L.R. Johnston, Ed. Chicago, Association of College and Research Libraries. 103-126.

Final model specification. n.d. Available: https://www.w3.org/community/ontolex/wiki/Final_Model_Specification [2017, December 27].

Flati, T., Moro, A., Matteis, L., Navigli, R. & Velardi, P. (2015). *Guidelines for linguistic linked data generation: multilingual dictionaries (Babelnet).* Available: https://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/ [2017, December 27].

Gillis-Webber, F. (2018). The construction of an ontological framework for bilingual lexicographic resources: applying linguistic linked data principles. M.Phil. dissertation. University of Cape Town.

Gouws, R.H. & Prinsloo, D.J. (2005). *Principles and practice of South African lexicography.* Stellenbosch: SUN MeDIA.

Gracia, J. (2017). Introduction to linked data for language resources [Practical session]. 2nd Summer Datathon on Linguistic Linked Open Data. 26 June.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). *Toward linked data-native dictionaries.* Available: https://elex.link/elex2017/wp-content/uploads/2017/09/paper33.pdf [2018, January 17].

Gracia, J. & Vila-Suero, D. (2015). *Guidelines for linguistic linked data generation: bilingual dictionaries.* Available: https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/ [2017, December 25].

Gracia, J., Villega, M., Gómez-Pérez, A. & Bel, N. n.d. *The Apertium bilingual dictionaries on the web of data.* Available: http://www.semantic-web-journal.net/system/files/swj1419.pdf [2017, December 31].

Heath, T. & Bizer, C. (2011). *Linked data: evolving the web into a global data space.* Morgan & Claypool Publishers.

Herbert, R.K. & Bailey, R. (2002). The Bantu languages: sociohistorical perspectives. In *Language in South Africa.* R. Mesthrie, Ed. Cambridge: Cambridge University Press. 50-78.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A. & Decker, S. (2012). An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web.* 14:14-44.

Hyvönen, E. (2012). *Publishing and using cultural heritage linked data on the semantic web.* Morgan & Claypool Publishers. DOI: 10.2200/S00452ED1V01Y201210WBE003

Keller, M.A., Persons, J., Glaser, H. & Calter, M. (2011). *Report on the Stanford Linked Data Workshop.* Available: https://www.clir.org/wp-content/uploads/sites/6/LinkedDataWorkshop.pdf [2017, December 26].

Labra Gayo, J.E., Kontokostas, D. & Auer, S. n.d. *Multilingual linked data patterns.* Available: http://www.semantic-web-journal.net/system/files/swj495.pdf [2017, December 27].

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the semantic web. *Language Resources & Evaluation.* 46:701-719.

McCrae, J.P. & Gracia, J. (2017). Introduction to the Ontolex-Lemon Model [Practical session]. 2nd Summer Datathon on Linguistic Linked Open Data. 26 June.

Ngcobo, M.N. Department of Linguistics and Modern Languages. (2010). *Only study guide for LIN3704: Language planning and language description.* Pretoria: University of South Africa.

Pretorius, L. (2014). The multilingual semantic web as virtual knowledge commons: the case of the under-resourced South African languages. In *Towards the multilingual semantic web.* P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 49-66.

PROV-Dictionary: Modeling provenance for dictionary data structures. (2013). Available: https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/ [2018, January 1].

PROV-O: The PROV ontology. (2013). Available: https://www.w3.org/TR/prov-o/ [2018, January 1].

Rafferty, P. (2016). Managing, searching and finding digital cultural objects: putting it in context. In *Managing digital cultural objects: analysis, discovery and retrieval*. A. Foster & P. Rafferty, Eds. London: Facet Publishing. 3-24.

Sachs, J. & Finin, T. (2010). *What does it mean for a URI to resolve?*. Available: http://ebiquity.umbc.edu/_file_directory_/papers/495.pdf [2017, December 26].

Simons, N. & Richardson, J. (2013). *New content in digital repositories: the changing research landscape*. Oxford: Chandos Publishing.

Statistics South Africa. (2012). *Census 2011 Census in brief*. Pretoria: Statistics South Africa. Available: http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf [2017, November 5].

Subfamily: Nguni (S.40). n.d. Available: http://glottolog.org/resource/languoid/id/ngun1276 [2018, February 11].

Tennis, J.T. (2007). Scheme versioning in the semantic web. In *Knitting the semantic web*. J. Greenberg & E. Méndez, Eds. 85-104.

Translating text. n.d. Available: https://cloud.google.com/translate/docs/translating-text [2018, January 5].

Van Erp, M. (2012). Reusing linguistic resources: Tasks and goals for a linked data approach. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 57-64.

Van Hooland, S. & Verborgh, R. (2014). *Linked data for libraries, archives and museums*. London: Facet Publishing.

Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J. & Aguado-de-Cea, G. (2014). Publishing linked data on the web: the multilingual dimension. In *Towards the Multilingual Semantic Web*. P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 101-117.

Villazón-Terrazas, B., Vilches-Blázquez, L.M., Corcho, O. & Gómez-Pérez, A. (2012). *Methodological guidelines for publishing government linked data*. Available: https://www.lri.fr/~hamdi/datalift/tuto_inspire_2012/Suggestedreadings/egovld.pdf [2017, December 25].

What is a language resource?. n.d. Available: http://www.elra.info/en/about/what-language-resource/ [2017, November 1].

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

Zgusta, L. (1971). *Manual of lexicography*. Prague: Academia, Publishing House of the Czechslovak Academy of Sciences.