

Article

# Conversion of the *English-Xhosa Dictionary for Nurses* to a Linguistic Linked Data Framework <sup>†</sup>

Frances Gillis-Webber

Library and Information Studies Centre, University of Cape Town, Woolsack Drive, Rondebosch, Cape Town 7701, South Africa; fran@fynbosch.com

<sup>†</sup> This paper is an extended version of the paper presented at the 6th Workshop on Linked Data in Linguistics (LDL-2018), 11th Edition of the Language Resources and Evaluation Conference (LREC 2018).

Received: 15 September 2018; Accepted: 2 November 2018; Published: 6 November 2018



**Abstract:** The *English-Xhosa Dictionary for Nurses* (EXDN) is a bilingual, unidirectional printed dictionary in the public domain, with English and isiXhosa as the language pair. By extending the digitisation efforts of EXDN from a human-readable digital object to a machine-readable state, using Resource Description Framework (RDF) as the data model, semantically interoperable structured data can be created, thus enabling EXDN's data to be reused, aggregated and integrated with other language resources, where it can serve as a potential aid in the development of future language resources for isiXhosa, an under-resourced language in South Africa. The methodological guidelines for the construction of a Linguistic Linked Data framework (LLDF) for a lexicographic resource, as applied to EXDN, are described, where an LLDF can be defined as a framework: (1) which describes data in RDF, (2) using a model designed for the representation of linguistic information, (3) which adheres to Linked Data principles, and (4) which supports versioning, allowing for change. The result is a bidirectional lexicographic resource, previously bounded and static, now unbounded and evolving, with the ability to extend to multilingualism.

**Keywords:** linguistic linked data framework; URIs; provenance; versioning; multilingualism; lexicography; linked data; resource description framework; RDF; ontolox-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 Xhosa-speaking nurses [1] (p. 1). It was published in 1935 in South Africa and is now in the public domain [1] (p. 1). EXDN is unidirectional, with English as the source language and Xhosa, an indigenous Bantu language from the Nguni language group (S40 in Guthrie's classification), the target language [1] (p. 1), [2] (p. 91), [3]. isiXhosa (referred hereon by its endonym) is one of eleven official languages in South Africa, where, with the exception of English, all are considered under-resourced languages due to the limited language resources available for each and the socio-economic constraints of the speakers [1] (p. 1), [4] (p. 72), [5] (pp. 49–53). Bantu languages are characterised by the use of a grammatical system to broadly categorise nouns, called a noun class system, with concordial agreement markers which show agreement between subject and verb, and between noun and modifier [6] (p. 429), [7] (p. 456). In the case of isiXhosa, there are fifteen noun classes, with affixes (morphemes, the smallest unit of a language) added to a word stem to create a word or phrase, rendering it an agglutinative language with a conjunctive orthography (the affixes and word stems are bound together when written) [6] (pp. 428–429, 433), [8] (p. 303).

Despite EXDN being published more than seventy-five years ago, as a language resource for an under-resourced indigenous African language, its content is still of value [1] (p. 1). EXDN is in print form, bounded and static, and by digitising the dictionary, it is converted from an analogue resource

into a simple digital resource, typically as images in JPEG format. However, by retrodigitising the artefact, converting it from a simple digital resource to a complex digital resource, it has the potential to become machine-interoperable. The form of complex digital objects can vary, with examples including: a collection of XML files; the same content but as HTML files with semantic markup to describe the content therein; as a dataset stored in a relational database or as Resource Description Framework (RDF). For the latter example, RDF “is a framework for representing information in the Web”, published by the World Wide Web Consortium (W3C), where data is described in short statements, called triples, which are of the form subject-predicate-object [9]. A subject can be an internationalised resource identifier (IRI) or a blank node; an object can be an IRI, blank node, or a literal; and a predicate can be an IRI only [9].

The lexical entry *Abdomen* can be expressed in the following short statements:

*Abdomen* is a lexical entry.

*Abdomen* is a word.

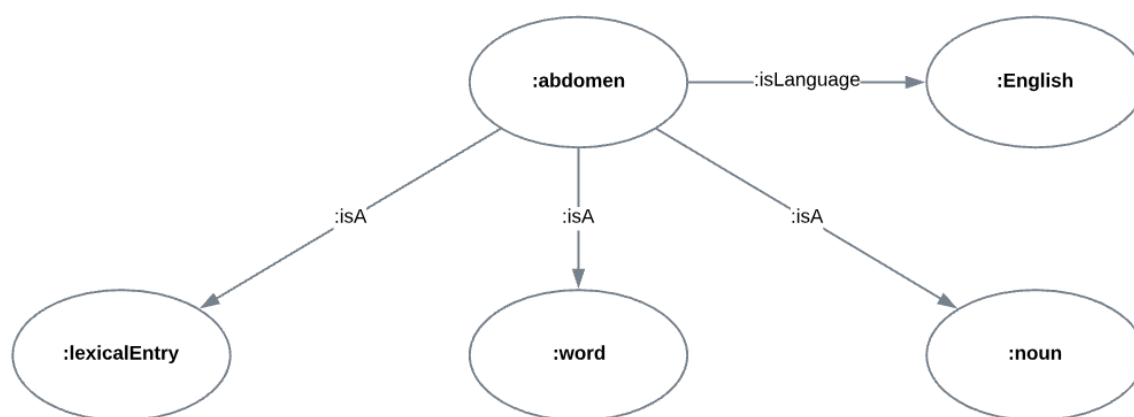
*Abdomen* is a noun.

*Abdomen* is an English term.

Using Turtle syntax, a human-readable serialisation of RDF, the same short statements can be described in RDF:

<i>Subject</i>	<i>Predicate</i>	<i>Object</i>
:abdomen	:isA	:lexicalEntry ;
	:isA	:word ;
	:isA	:noun ;
	:isLanguage	:English .

These short statements are visualised in Figure 1, using node-edge-node structure.



**Figure 1.** Visualisation of the lexical entry *Abdomen*, described using RDF triples.

In a 2009 TED talk, Tim Berners-Lee said that by creating relationships in data, “the more things you have to connect together, the more powerful it is”; to create these relationships, he recommends putting data on the web and using Uniform Resource Identifiers (URIs), where the URI is the same as an IRI, except the former only allows for ASCII characters [9,10]. In the context of this paper, URIs are used. URIs differ from Uniform Resource Locators (URLs) conceptually, where the latter refers to the location of a document, but the former serves to identify: not only documents, but objects or concepts as well [11], [12] (p. 25), [13] (p. 46). Returning to the short statements serialised in Turtle, the “:English” object can be replaced with a URI from an external resource, for which Lexvo.org has been selected. Another triple has been added, where the denotation of the lexical entry *Abdomen* is identified as the DBpedia resource “Abdomen”. The triples are now as follows:

Subject	Predicate	Object
:abdomen	:isA	:lexicalEntry ;
	:isA	:word ;
	:isA	:noun ;
	:isLanguage	http://lexvo.org/id/iso639-3/eng ;
	:isDenotedBy	http://dbpedia.org/resource/Abdomen .

If one had to describe a second lexical entry, *Isisu*, in RDF triples:

:isisu	:isA	:lexicalEntry ;
	:isA	:word ;
	:isA	:noun ;
	:isLanguage	http://lexvo.org/id/iso639-3/xho ;
	:isDenotedBy	http://dbpedia.org/resource/Abdomen .

Because both lexical entries share the same denotation, equivalence between *Abdomen* and *Isisu* can be inferred. These short statements are visualised in Figure 2.

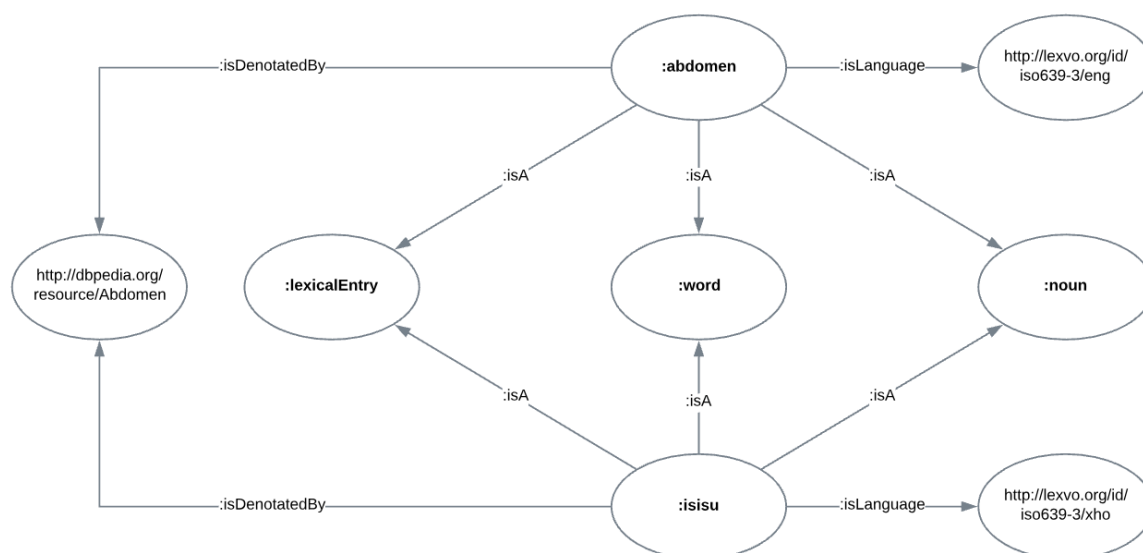


Figure 2. Visualisation of the lexical entries *Abdomen* and *Isisu*.

Each subject, predicate or object can be converted to a URI (unless it is a literal or a blank node), where a URI can be defined relative to the resource or it can be from an external data source. Linked Data can thus be defined as the techniques and best practices for publishing structured data on the web, for which describing the data in RDF, and identifying and creating links between the data are fundamental components thereof [13] (p. 3), [14] (p. 4).

Berners-Lee has identified the following Linked Data principles [11]:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

By extending the digitisation efforts of EXDN, a little-known lexicographic resource, from a human-readable digital object to a machine-readable state, using RDF and described according to Linked Data principles, semantically interoperable structured data can be created, thus enabling

EXDN's data to become “shareable, extensible, and easily reusable” [15], and in the words of Berners-Lee, able to be “combined into something more interesting than the original” [10].

In the case of EXDN, in print form, it is a bilingual, synchronic and unidirectional resource. However, by converting the lexical entries to a Linguistic Linked Data framework (LLDF), where in the context of this paper, an LLDF can be defined as a framework:

1. which describes data in RDF,
2. using a model designed for the representation of linguistic information,
3. which adheres to Linked Data principles, and
4. which supports versioning, allowing for change;

not only does the resource become bidirectional and diachronic, with the opportunity to extend to multilingualism, but it also allows for the “aggregation and integration of linguistic resources” [15], thereby serving as a potential aid in the development of future language resources for isiXhosa, an under-resourced language in South Africa [1] (p. 1).

Examples of projects which serve their data as Linked Data include Princeton WordNet 3.1 (PWN), a large lexical database, DBpedia, a knowledge base which extracts structured content from various Wikimedia projects, BabelNet, a multilingual encyclopedic dictionary, and the Apertium Bilingual Dictionaries (ABD), with the latter three projects running on a Virtuoso server [16–18]. Although not the primary focus, the conversion of EXDN to an LLDF was also intended as a proof of concept to extend the human-readable view of an online dictionary to a machine-readable view using a LAMP stack (Linux, Apache, MySQL and PHP), on a shared web hosting platform.

The rest of the paper is organised as follows: in Section 2, the process of digitising EXDN is briefly described; in Section 3, the methodological guidelines for the construction of an LLDF, as applied to EXDN, is considered; the conclusions of the paper, as well as future work, are presented in Section 4.

Note: namespaces are presumed defined for all code examples. See Appendix A for the list of defined namespaces.

## 2. The Digitisation of EXDN

The workflow to digitise EXDN is illustrated in Figure 3.

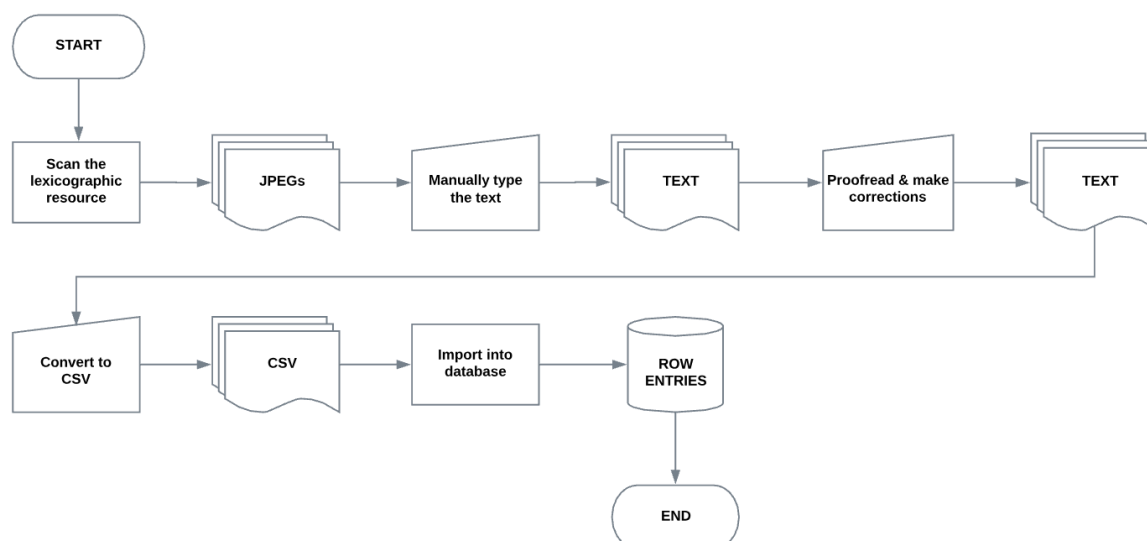


Figure 3. Workflow illustrating the digitisation process.

Digitisation of the resource was attempted using a high-resolution scanner and optical character recognition (OCR) in Adobe Acrobat Pro. However, the text could not be recognised, and as a result, the resource was manually typed. 1748 lexical entries were imported into the database, and at least 200 of those entries were estimated to be translation equivalents due to the restricted treatment of the lemma sign, for which the “source language item, represented by the lemma sign, is co-ordinated with a single target language item” [19] (p. 154). Some lexical entries, although outdated, will be published and indicated as such; an example is the lexical entry with the lemma *Sanatorium*. However, lexical entries which no longer have relevance, for example, the article *Benger’s Food*, a medicinal tinned food produced until the middle of the 20th century, have been excluded from publication [20,21].

A Dictionary Writing System (DWS) was developed by the author, using PHP as the web scripting language and a MySQL database. The purpose of the DWS is to manage the languages, lexical entries, and lexical concepts of the project; to prepare the lexical entries, lexical concepts, and the lexicons for publication; and to generate the different formats required: RDF serialised using Turtle and N-Triples. The publishing process of a single lexical entry is illustrated in Figure 4.

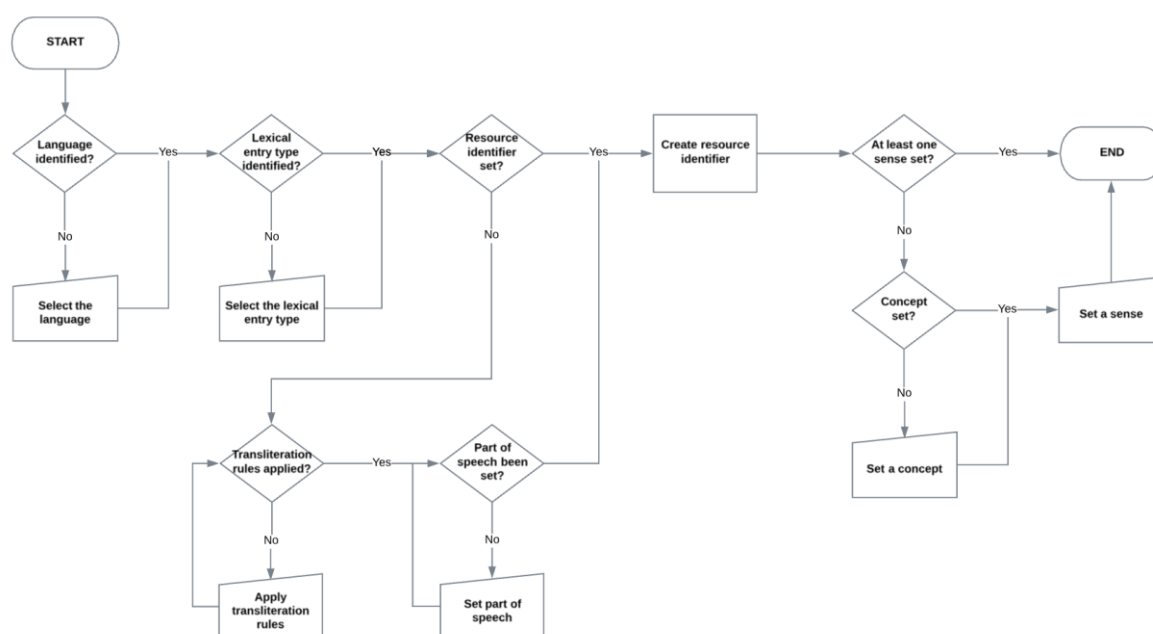


Figure 4. Illustration of the publication process of a single lexical entry.


Each lexical entry must have at least one sense, and once a sense has been set, the lexical entry is automatically published the following day, this includes lexical concepts (published either as a new concept or as an existing concept updated to include the lexicalised sense of the lexical entry), and the lexicon (as an update, showing the lexical entry as a new member of the lexicon). Figure 5 shows a selection of lexical entries in the DWS for the article stretch A, with *en-n-abdomen* scheduled for (re)publication.

**Lexicons**

Language

Status

Search by lemma

View by letter **A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | **

a pint, redirects to octarius	<a href="#">view entry</a>
en-n-abdomen	<a href="#">view entry</a>
<b>SCHEDULED FOR REPUBLICATION</b>	
abdominal	<a href="#">view entry</a>
Outstanding: Identifier, Sense	
en-n-abdominal_operation	<a href="#">view entry</a>
Outstanding: Sense	
abnormality	<a href="#">view entry</a>
Outstanding: Identifier, Sense	
Can't add DBpedia: need to disambiguate. Confirm Xhosa meaning before making a decision.	
en-n-abortion	<a href="#">view entry</a>
Outstanding: Sense	
Definition is possibly a lexical equivalent	

Figure 5. Selection of lexical entries in the DWS.

### 3. Methodological Guidelines for the Construction of a Linguistic Linked Data Framework

In 2011, Villazón-Terrazas et al. proposed methodological guidelines for publishing Government Linked Data, with the guidelines taking an iterative approach [22] (pp. 4–13). In 2014, Vila Suero et al. proposed methodological guidelines based on Villazón-Terrazas et al.'s iterative model, but these included the language aspect [23] (pp. 103–115). Briefly described here, Vila Suero et al.'s guidelines are (1) *Specification*, (2) *Modelling*, (3) *Generation*, (4) *Linking*, and (5) *Publication*, where *Specification* refers to the identification and analysis of the data sources, and URI design; *Modelling* refers to the identification and creation of domain vocabularies to use, as well as ontology localisation; *Generation* refers to the transformation of the data sources to RDF, the identification of the languages used, and the consideration of encoding issues; *Linking* refers to interlinking with external resources; and *Publication* refers to the publication of the dataset and its metadata [23] (pp. 103–115).

In 2015, Gracia and Vila-Suero published guidelines for generating Linguistic Linked Data for multilingual dictionaries and other lexical resources, and they are broadly described here: (1) identify the model, (2) select the vocabularies, (3) analyse the data source(s), (4) model the source lexicon, the target lexicon, and the translation set, (5) model a lexical entry, (6) design the URIs, (7) transform the data into RDF, (8) publish the RDF dataset, and (9) publish the metadata [24].

The case study as a research method was adopted for the project, using EXDN as a single case design [25] (pp. 1–2). The methodological guidelines identified by Vila Suero et al. and Gracia and Vila-Suero were used to test the construction of an LLDF, however, as EXDN is a bilingual resource only, cognisance was taken of an additional use case: namely that the LLDF should allow for extensibility to multilingualism, should the lexical data necessitate it.

Using aspects of Vila Suero et al.'s methodological guidelines and Gracia and Vila-Suero's guidelines for publishing Linguistic Linked Data, the following methodology was identified when converting EXDN to a Linguistic Linked Data framework:

- Step 1: Identify the use cases
- Step 2: Select the model with which to describe the language data in RDF
- Step 3: Identify the external resources to link to
- Step 4: Identify the versioning strategy

- Step 5: Identify the RDF formats
- Step 6: Identify the URI strategy
- Step 7: Consider the lemmatisation approach
- Step 8: Model lexical entries, a lexicon, and a lexical concept
- Step 9: Generate the RDF data
- Step 10: Publish the RDF data

Each step is expanded upon in more detail.

### **Step 1: Identify the use cases**

Due to the size of the dataset, it was not possible to manually convert every lexical entry to RDF, instead key characteristics of the lexical entries were abstracted, with these abstractions serving as use cases that the selected model would have to support [13] (p. 12). The following use cases were identified:

- M1:** Modelling a lexical entry that offers a restricted treatment of the lemma sign.
- M2:** Modelling a lexical entry that offers a paraphrase of meaning of the lemma sign.
- M3:** Modelling a lexical entry that contains a cross-reference entry.
- M4:** Modelling a lexical entry that offers a comment on semantics.
- M5:** Modelling a plural form for an African language in the lexical entry.
- M6:** Modelling a lexical entry with a stem as the lemma.
- M7:** Modelling a lexical entry with a derived noun as the lemma.
- M8:** Modelling a lexical entry for a derived noun, with the plural form as the lemma.
- M9:** Modelling a translation relation between a source and target sense, which do not share the same lemmatisation approach.
- M10:** Modelling a lexical entry which has an outdated sense.

### **Step 2: Select the model with which to describe the linguistic data in RDF**

In 2017, the 2nd Summer Datathon on Linguistic Linked Open Data was held in Spain; described as “unique in its topic worldwide”, it is a biennial datathon series focusing on the field of Linked Data as applied to Linguistics, and one of its aims was to show how to migrate linguistic data from existing data sources, publishing it online as Linked Data [26]. At this datathon, the model presented for describing Linguistic Linked Data was Ontolex-Lemon.

Ontolex-Lemon’s predecessor, known as *lemon*—the Lexicon Model for Ontologies, was developed by the Monnet project from 2010, and in May 2016 it was published as a W3C vocabulary under the name of Ontolex-Lemon [27,28]. At time of writing, the model is actively maintained, undergoing continuous development by the W3C Ontology-Lexica Community Group [29]. Ontolex-Lemon (and *lemon*) represents lexicons and machine-readable dictionaries relative to ontologies, described as the *ontology-lexicon interface*, with the ontology forming a “shared conceptualisation” and the lexicon describing the “lexical encoding of that conceptualisation in words”, using a principle called *semantics by reference*, where “the meaning of a word is given by reference to an ontology, resulting in a clean separation between the lexical and semantic layer” [30] (p. 16). The *lemon* model was influenced by Lexical Markup Framework (LMF), as well as the models: LexInfo, Linguistic Information Repository (LIR), the Linguistic Meta Model (LMM), the semiotics.owl ontology design pattern, and the Senso Comune core model [28].

LMF, designed between 2003 and 2008, is ISO standard 24613:2008 and it provides a standardised framework for natural language processing (NLP) and machine-readable dictionaries [31]. While it is able to represent linguistic information, it cannot represent lexicons to ontologies [32], [33] (p. 3), and although it describes itself as interoperable, LMF has been criticised for its inability to establish interoperability between different lexicons, as well as its vagueness for use when applied to different



contexts [30] (p. 27), [34] (p. 96). Despite this, *lemon* (and thus Ontolex-Lemon) drew heavy inspiration from LMF, with LMF's core ontology adopted by *lemon*, as well as classes and entities imported from LMF, however, in order to describe the ontology-lexicon interface, additional vocabulary was added [28].

LexInfo proposed a model that unified LMF with OWL, the Web Ontology Language by W3C, with it building conceptually on three components: the models LingInfo, LexOnto, and LMF [30] (p. 27), [35] (p. 30). LingInfo provides a mechanism “for modelling label-internal linguistic structure”, such as inflection, interpreted as terms; LexOnto enables “the representation of label-external linguistic structure”, mapping to ontological structures [35] (p. 30). Both models are complementary, and by combining aspects of these models within LexInfo, linguistic information could be associated with ontology elements in a way that was reusable across systems [35] (pp. 29–30). Although RDF, RDF Schema (RDFS), OWL, and Simple Knowledge Organization System (SKOS) can associate labels with ontology elements, the linguistic information thereof is not able to be described; however, SKOS does allow for further typology, for example, identifying a label as “preferred” [23] (p. 110), [35] (pp. 29–30).

LIR focuses on the variations of terms (such as acronyms and transliterations), where it explicitly defines translation relations between term variants, using an OWL meta-ontology which can be associated with any element of an OWL ontology [36] (p. 106), [37] (p. 822).

Figure 6 shows a timeline of the models under discussion (ending December 2017), in terms of active development for each. Four models were ultimately reviewed: LMF, LexInfo, LIR, and Ontolex-Lemon. The original *lemon* model was not considered due to development being undertaken by the W3C Ontology-Lexica Community Group, with Ontolex-Lemon as the result [27,28].

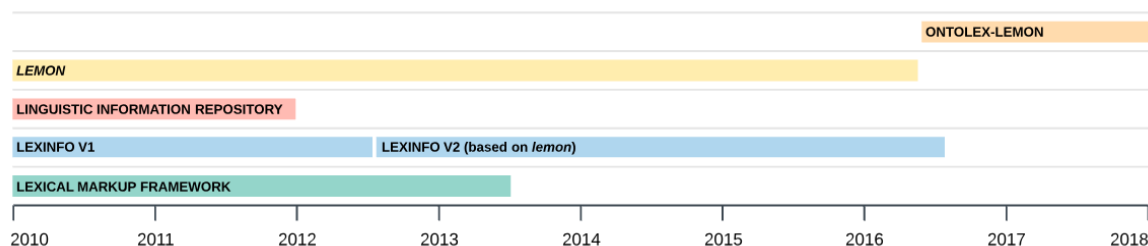


Figure 6. Timeline of the models in development.

Considering the use cases identified in Step 1, the requirements of the model are broadly defined in sections based on those identified by Cimiano et al. [35] (pp. 29–33), where the model should allow for:

1. **Interoperability;**
2. **Separation and independence:** where there is separation between the lexical and the ontological layer, with linguistic information able to be modelled separately;
3. **Linguistic information:** where structured linguistic information can be captured;
4. **Morphological decomposition:** necessary when working with an agglutinative language such as isiXhosa;
5. **Multilinguality:** where there is support for multilingualism and translation relations, beyond language tagging;
6. **Ontological representation:** where meaning can be represented by an external ontology entity (referred to as “arbitrary ontologies” in Cimiano et al.) [11] (p. 33);
7. **Linked Data principles:** where there is adherence to the principles of Linked Data, listed in Section 1.

Using Table 1 from Cimiano et al. [35] (p. 33), Table 1 was updated to include the requirements as broadly defined above, with SKOS included for informational purposes.

As shown in Table 1, Ontolex-Lemon was the only model which fulfilled all the modelling requirements.



**Table 1.** Features of the models which correspond to the modelling requirements, derived from Cimiano et al. [35] (p. 33).

	1 Inter- operability	2 Separation & Independence	3 Linguistic Information	4 Morphological Decomposition	5 Multi- linguality	6 Ontological Representation	7 Linked Data Principles
SKOS	Yes	No	No	No	No	No *	Yes
LMF	No	No	Yes	Yes	Yes	No	No
LexInfo	Yes	Yes	Yes	Yes	No	Yes	Yes
LIR	Yes	Yes	Yes	No	Yes	Yes	Yes
Ontolex-Lemon	Yes	Yes	Yes	Yes	Yes	Yes	Yes

\* In Cimiano et al., this is indicated as “Not applicable” [35] (p. 33).

PWN and BabelNet have published their datasets in RDF format using the *lemon* model, as has Zhishi.lemon, the lexical realisation of Zhishi.me, a Chinese dataset in the Linked Open Data cloud, and the ABD, a machine translation platform with up to 40 language pairs [18] (p. 2), [38] (p. 47). Exemplars of other language resources converted to Linked Data using Ontolex-Lemon (or its predecessor, *lemon*) include “‘Al-Qāmūs Al-Muḥit”, a Classical Arabic dictionary, the Pattern Dictionary of English Verbs, K Dictionary Series’ German monolingual dictionary, and an exemplar dictionary article from “Dictionnaire étymologique de l’ancien français” [39] (p. 325), [40] (pp. 590–591), [41] (p. 2). These resources validate the use of the model by virtue of precedence; Ontolex-Lemon has thus been selected and it serves as the basis for the remaining steps.

Ontolex-Lemon consists of a core module in which the primary element is the *Lexical Entry*, where a word, multiword expression, or an affix can be represented [28]. Senses can be defined for a lexical entry, with the meaning thereof a reference to an ontology entity [28]. Additional modules include Syntax and Semantics (*synsem*), Decomposition (*decomp*), Variation and Translation (*vartrans*), and Metadata (*lime*) [28].

### Step 3: Identify the external resources to link to

The following ontologies and vocabularies were identified for use:

- **DBpedia** (a cross-domain ontology used to identify resources) [16];
- **Dublin Core Metadata Initiative** (used to describe the properties of resources) [42];
- **FOAF** (used to describe properties and identify resources) [43];
- **Library of Congress Name Authority File** (controlled vocabulary used to identify persons and organisations) [44];
- **Library of Congress Subject Headings** (controlled vocabulary used to categorise resources) [45];
- **LexInfo** (used to represent lexical information) [46];
- **Medical Subject Headings** (MeSH) (controlled vocabulary used to identify and categorise resources in the medical domain) [47];
- **Multilingual Morpheme Ontology** (MMoOn) (a multilingual morpheme ontology used to express linguistic concepts and relations) [48];
- **PROV Ontology** (PROV-O) (used to represent provenance information) [49];
- **Princeton WordNet 3.1** (the RDF interface used for Princeton WordNet) [50]; and
- **VOID Vocabulary** (a vocabulary for expressing metadata about datasets) [51].

### Step 4: Identify the versioning strategy

According to Di Maio, knowledge is “partial/incomplete/imperfect, with very few exceptions” [52]. 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; a powerful notion for under-resourced languages [1] (p. 5), [10,53]. According to Bouda and

Cysouw, 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 [54]. Like ontologies, RDF datasets are not static, and they too evolve over time [1] (p. 5), [12] (p. 94). 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 [1] (p. 5), [12] (p. 95). As RDF has an open-world assumption, the conversion of each lexical entry to RDF can never be regarded as fully complete; instead, completeness is considered to be on a continuum [13] (pp. 61, 161). This “incompleteness” is mitigated by focussing on the provenance and versioning of an LLDF whose instances are constantly evolving.

Versioning can be distinguished between ontologies and RDF datasets. When discussing ontology versioning, Klein and Fensel identified possible scenarios of an ontology after it undergoes a change [55] (p. 7):

1. The ontology changed invisibly, that is, there was no notification of the change, prior or post the event. From this, one scenario can result:
  - a. A new version of the ontology replaces a previous version. Any previous versions are no longer available.
2. The ontology changed visibly, that is, there was a notification of the change, prior or post the event. From this, several scenarios can result:
  - a. A new version of the ontology replaces a previous version. Any previous versions are no longer accessible.
  - b. A new version of the ontology replaces a previous version. Any previous versions remain accessible.
  - c. A new version of the ontology replaces a previous version. Any previous versions remain accessible. There is also an explicit specification of the changes between the previous version and the new version.

Moving onto the versioning of RDF datasets, within the context of Linguistic Linked Data, 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” [1] (p. 6), [56]. When the generation and publication of RDF data for the ABD was detailed by Gracia et al., versioning was not included in the discussion [18]. Although briefly mentioned by McCrae et al., Gracia et al., Eckart et al., van Erp, and De Rooij et al. [18,53,57–59], 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” [1] (p. 6), [56].

When describing the generation of RDF for the ABD, Gracia et al. talked of three RDF files: one per lexicon, and the third for the translations [18]. 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 [1] (p. 6). However, in the context of EXDN, it was felt that publishing only at the lexicon-level could become unmanageable over time, and instead it would be more practical to implement versioning at the lexical entry-level as well [1] (p. 6). 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 [1] (p. 6). For each version of a lexical entry, the file contains all information pertaining to the lexical entry, its senses, and translation relations for which any of its senses is the source [1] (p. 6).

The following components were thus identified for the versioning:

- versioned URIs for lexical entries, lexicons, and lexical concepts,

- provenance metadata to describe the versions, with the latest version mapping to previous versions [57], and
- the generation of files, one for each version of the lexical entries, lexical concepts, and lexicons, with each lexical entry, lexical concept, and lexicon treated as an individual repository, containing a set of statements [1] (p. 6), [60] (p. 2).

Using the ontology versioning scenarios by Klein and Fensel as a guide, for each individual repository:

- the change should be visible,
- previous versions should remain accessible, and
- a changelog between versions should be explicitly specified.

For each URI:

- it should be persistent,
- it should not be deleted,
- however, it can be deprecated or superseded [61] (p. 4).

### **Step 5: Identify the RDF formats**

Within the context of EXDN, URIs are dereferenceable, and the machine-readable view for URIs of lexical entries, lexical concepts and lexicons return RDF serialised in Turtle (where Turtle is human-friendly, used both for modelling and for its readability) [12] (p. 22), [62] (p. 19). The dataset will be periodically uploaded as a data dump to different data repositories, and for this, the serialisation will be N-Triples (where each triple is written one per line), with it described as the “best for huge data sets” [17].

### **Step 6: Identify the URI strategy**

A URI has been defined by Archer, Goedertier and Loutas as “a compact sequence of characters that identifies an abstract or physical resource” where it serves as “a locator, a name, or both” [63]. The following principles have been identified for URIs:

- URIs should be short, stable and persistent [14,63,64];
- they should be HTTP(S) URIs [11,64];
- URIs should be dereferenceable, returning a representation that is human- and machine-readable [12,62];
- URIs should distinguish between the resource, and the document describing that resource [13,62];
- a URI’s identifier portion should be unique and unambiguous [65,66];
- if it is necessary to avoid language bias, where a URI’s identifier portion can rather be presented as agnostic of any language, then opaque URIs should be used, with the identifier typically represented by a number [23] (p. 107), [67] (p. 5);
- if a URI is expected to be looked up by both a web browser for human consumption and a software agent, then a URI’s identifier portion should be descriptive, as this is more “user-friendly” and “meaningful” for the human user [23] (p. 26), [63] (p. 18), [22] (p. 6).

In the subsections that follow, fragment identifiers and URI patterns are discussed in detail; concluding with a brief discussion of resource identifiers within the context of EXDN.

#### **3.1. Fragment Identifiers**

Fragment identifiers are an optional part of the URI, positioned at the end, and are of the pattern “#example” [1] (p. 2). Although the usage of fragment identifiers has 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” [1] (p. 2), [14] (p. 31). 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) [1] (p. 2).

According to Sachs and Finin, the URI should resolve “not to the address, but to all known information about the resource” [68]; from this one can infer that when information for a sub-resource is returned, then information for the parent resource should also be returned [1] (p. 2). Conversely, when information for a parent resource is returned, information of any sub-resources should also be returned [1] (p. 2). By using fragment identifiers, 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 [1] (p. 2). 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 [1] (p. 2).

### 3.2. The URI Pattern

When working with the EXDN data, the following URI use cases were identified [1] (p. 2):

- U1: A URI that identifies a resource
- U2: A URI that identifies a sub-resource in relation to the parent resource
- U3: A URI that identifies a version of the resource
- U4: A URI that identifies a version combined with a sub-resource
- U5: A URI that identifies a document describing the resource in U1
- U6: A URI that identifies a document describing the resource in U3

Archer, Goedertier and Loutas have recommended a pattern for URIs:

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

where:

- {domain} is the host,
- {type} is the resource being identified,
- {concept} refers to a real world object or a collection, and
- {reference} is the local reference for the resource being identified [1] (p. 2), [63] (p. 19).

Gracia and Vila-Suero used the pattern recommended by Archer et al. in their methodological guidelines for language resources published in 2015 [24], with an example lexical entry for “bench” as follows [1] (p. 2):

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

Within the context of EXDN, it was not necessary to identify the collection using {concept} so this portion of Archer et al.’s pattern was excluded. A requirement of Ontolex-Lemon (and previously *lemon*) is that a lexicon only contains lexical entries of the same language, with Gracia and Vila-Suero identifying the language of the lexicon in their URIs. Within the context of EXDN, this too was excluded as it was deemed preferable that the URI pattern remains agnostic of the model; should the model be replaced, the persistence and longevity of existing URIs would not be impacted by this change.

For each of the six URI use cases identified previously, this simplified pattern has been applied, 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 [1] (p. 3, 4).

A URI which identifies a resource has the form:

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

where:

- {http(s):} is the http: or https: scheme
- {Base URI} is the namespace
- {Resource Path} is, for example, *entry* for a lexical entry, and *lexicon* for a lexicon
- {Resource ID} is the resource identifier

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:

**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:

**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:

**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 [63] (p. 6).

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

**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.

Example URIs 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:

**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 descriptive approach for identifying lexical resources in E1 (“bench-n-en”) was similarly adopted for EXDN lexical entries, however the elements were reordered to aid programmatic extraction (should it be necessary) [1] (p. 4):

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

where:

- {Language Code} is the lowercase form of an ISO 639 code
- {POS} is an abbreviated form of the part-of-speech
- {Lemma} is the lowercase form of the lemma, with diacritics removed and hyphens or spaces replaced with underscores

The descriptive approach was also used for identifying lexicons, where the resource identifier is a lowercase form of an ISO 639 code. For lexical concepts, the opaque approach was used, where the resource identifier is an incremental number.

The URI strategy requires patterns to be identified for both the URIs and the resource identifiers. Depending on the type of resource being identified, a different pattern may need to be employed for resource identifiers: using descriptive or opaque identifiers, where the former is “human readable” [64] (p. 25) and “user-friendly” [63] (p. 18), but the latter avoids language bias by being language-agnostic [23] (p. 107). Within the context of EXDN, descriptive resource identifiers were employed for lexical entries and lexicons, where both are specific to a single language; for lexical concepts, which can be multilingual, opaque resource identifiers were used [1] (p. 4).

#### Step 7: Consider the lemmatisation approach

A lemma is the address of a lexical entry which is used to retrieve information; a word, a word stem, or a multiword expression can all be lemmas in the same lemma list [31] (pp. 64–67). The lexicographic tradition for the lemmatisation of nouns and verbs, namely word versus stem, will vary depending on the conjunctiveness or disjunctiveness of the orthography of the language concerned [31] (p. 68). Due to this variation, the lemmatisation approach for isiXhosa, an agglutinative language, had to be considered when constructing the LLDF [19] (pp. 75–84).

As EXDN is unidirectional with English as the source language, the word approach was employed for lemmatisation. However, by converting the data to RDF using Ontolex-Lemon, for lexical entries where there is full equivalence, these target language items had to be created as lexical entries as well, with the EXDN data thus becoming bidirectional.

Lexicographic resources available online are not constrained by physical space [19] (p. 76), and a hybrid approach could be considered for EXDN. For English lexical entries, the word approach was used; for isiXhosa lexical entries, the word approach was selected for nouns and the stem approach selected for verbs [19] (pp. 79–80), [69] (pp. 760–761), [70] (p. 168). However, if there are frequently used forms for verbs, then it is anticipated that the word approach may also be used for these forms [71] (p. 32).

#### Step 8: Model lexical entries, a lexicon, and a lexical concept

When modelling a lexical entry in RDF, the following were identified as requirements:

- a description of the lexical entry, linking to the external resources identified in Step 3;
- metadata of the lexical entry;
- provenance information;
- a brief description of any related resources;
- a description of the lexicon which contains the lexical entry;



- and a description of the document which describes the lexical entry.

Each of the use cases identified in Step 1 were modelled, using the RDF format identified in Step 5, and taking into account the lemmatisation approach identified in Step 7.

When modelling a lexicon, similar requirements were identified, however, due to the potential size of the lexicon (where Ontolex-Lemon requires one lexicon per language), information of its related resources, namely lexical entries, were not included.

A lexical concept represents “a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses” [28], and within the context of EXDN, it can also serve as a shared conceptualisation between two or more senses from different languages for which there is full equivalence. The modelling requirements of a lexical concept are not dissimilar to that of a lexical entry, except that a lexical concept models a synset relation using *dct:references*, where the lexical concept is expressed by a member of the given synset [72] (p. 665). A sense of a lexical entry shares this relation by way of *ontolex:lexicalizedSense*, where the canonical form of the lexical entry is the same as the member of the given synset. The implication of this for EXDN is that each language represented within a lexical concept will need to have a relation declared to a synset of that language. For English senses, this is possible using PWN, however for isiXhosa senses, as an equivalent WordNet is not available, a URI has been created but it serves to identify only and is not dereferenceable. A lexical concept is modelled thus:

```

1      :000000001
2      a          skos:Concept , ontolex:LexicalConcept , prov:Entity ;
3      ontolex:lexicalizedSense :entry/en-n-abdomen#sense1 ;
4      ontolex:lexicalizedSense :entry/xh-n-isisu#sense1 ;
5      owl:sameAs      mesh:M000005 ;
6      dct:subject      mesh:D000005 ;
7      ontolex:isConceptOf      dbr:Abdomen ;
8      dct:references      pwn:05564576-n#abdomen-n ,
                           <https://wn.londisizwe.org/xh/000000001-n#isisu> .

```

The versioning of a lexicon is demonstrated in the section that follows so the modelling of a lexicon is not covered here. For the modelling of lexical entries, dictionary articles were identified that could serve as exemplars for the use cases (M1–M10) identified in Step 1. As a use case was modelled, the DWS was updated so that each of the requirements identified in the modelling could be managed via the DWS, such as setting the singular and plural forms of a lexical entry.

*Modelling the article: “Breath. Umphefumlo.” [73]*

The following points were identified for modelling:

- *Umphefumlo* is a translation equivalent
- *Umphefumlo* is a derived noun
- The stem is: -phefumlo
- The plural of *breath* is *breaths*
- The plural of *umphefumlo* (isiXhosa Noun Class 7) is *imiphefumlo* (isiXhosa Noun Class 8)

From this, the following resources were identified that would need to be modelled in RDF:

- Lexical entry: *en-n-breath* (of type Word)
- Lexical entry: *xh-n-umphefumlo* (of type Word)
- Lexical entry: *xh-n-phefumlo* (of type Stem)
- Lexical entry: *xh-n-um* (of type Affix)
- Lexical entry: *xh-n-imi* (of type Affix)



- Lexical concept: shared conceptualisation for *en-n-breath*, *xh-n-umphefumlo* and *xh-n-phefumlo*

If the stem approach had been selected for the lemmatisation of isiXhosa nouns, then only *xh-n-phefumlo* and *en-n-breath* would be described in RDF. Because the word approach has been selected, the additional forms, as listed above, are also required.

The description of the lexical entry *xh-n-umphefumlo* is modelled thus:

```

1   :xh-n-umphefumlo
2       a          ontolex:LexicalEntry , ontolex:Word , mmoon:DerivedNoun ;
3       lexinfo:partOfSpeech  lexinfo:Noun ;
4       dct:language  <http://id.loc.gov/vocabulary/iso639-2/xho> ,
                        <http://lexvo.org/id/iso639-1/xh> ;
5       mmoon:consistsOfStem   :xh-n-phefumlo ;
6       rdfs:label             "umphefumlo"@xh ;
7       ontolex:canonicalForm  :xh-n-umphefumlo#lemma ;
8       ontolex:lexicalForm    :xh-n-umphefumlo#singular ,
                                :xh-n-umphefumlo#plural ;
9       ontolex:sense          :xh-n-umphefumlo#sense1 ;
10      ontolex:evokes <https://londisizwe.org/concept/000000000> .
11
12 :xh-n-umphefumlo#lemma
13     a          ontolex:Form ;
14     ontolex:writtenRep "umphefumlo"@xh .
15
16 :xh-n-umphefumlo#singular
17     a          ontolex:Form ;
18     ontolex:writtenRep "umphefumlo"@xh ;
19     lexinfo:number lexinfo:singular ;
20     mmoon:consistsOfAffix :xh-n-um ;
21     mmoon:consistsOfStem :xh-n-phefumlo ;
22     rdf:_1           :xh-n-um ;
23     rdf:_2           :xh-n-phefumlo ;
24     lonvoc:inNounClass lonvoc:IsiXhosaNC7 .
25
26 :xh-n-umphefumlo#plural
27     a          ontolex:Form ;
28     ontolex:writtenRep "imiphefumlo"@xh ;
29     lexinfo:number  lexinfo:plural ;
30     mmoon:consistsOfAffix :xh-n-imi ;
31     mmoon:consistsOfStem :xh-n-phefumlo ;
32     rdf:_1           :xh-n-imi ;
33     rdf:_2           :xh-n-phefumlo ;
34     lonvoc:inNounClass lonvoc:IsiXhosaNC8 .
35
36 :xh-n-umphefumlo#sense1
37     a          ontolex:LexicalSense ;
38     ontolex:isLexicalizedSenseOf
        <https://londisizwe.org/concept/000000000> .

```

where:

- Line 2: indicates that it is a derived noun

- Line 5: indicates the lexical entry of the stem *xh-n-phefumlo*
- Lines 16–24: indicate the singular form
- Lines 20–21: indicate the affix and stem of this form
- Lines 22–23: indicate the order in which the derived noun is composed
- Line 24: indicates the isiXhosa noun class to which this form belongs
- Lines 26–34: indicate the plural form
- Lines 30–31: indicate the affix (note the difference to the singular form) and stem of this form
- Line 34: indicates the isiXhosa noun class to which this form belongs (note the difference to the singular form)

With the modelling of the lexical entry *xh-n-umphefumlo*, the following use cases have been addressed:

**M1:** Modelling a lexical entry that offers a restricted treatment of the lemma sign.

**M5:** Modelling a plural form for an African language in the lexical entry.

**M7:** Modelling a lexical entry with a derived noun as the lemma.

**M9:** Modelling a translation relation between a source and target sense, which do not share the same lemmatisation approach.

For the use case **M9** (and **M1**), as a sense is identified as a *ontolex:isLexicalizedSenseOf* a concept, senses from other lexical entries, be they words, derived words or stems, which are lexicalised to the same lexical concept, are equivalents.

The MMoOn ontology has been essential when modelling the isiXhosa lexical entries. When modelling lexical entries using Ontolex-Lemon, only the subclasses Word, MultiWord Expression and Affix are available. A new module for morphology within Ontolex-Lemon, based on the MMoOn ontology, is anticipated so although a stem cannot currently be modelled as a subclass of *ontolex:LexicalEntry*, this may be a future possibility [74].

Although not described in detail here, all the use cases identified in Step 1 could be modelled using Ontolex-Lemon. However, if the selected model was not able to support the modelling requirements of the use cases, then Step 2 would need to be revisited.

### Step 9: Generate the RDF data

When converting data to RDF, three approaches can be taken: (1) automatic conversion, (2) partial scripted conversion, and (3) modelling, which is then followed by scripted conversion [72]. For EXDN, the third approach was adopted, according to the versioning strategy identified in Step 4, using the modelling from Step 8. The identification and selection of external resources for each lexical entry and lexical concept to link to (notably DBpedia and MeSH) is manually managed in the DWS and automatically included during generation.

In the subsections that follow, provenance and versioning, used in conjunction with versioned URIs, are discussed in more detail.

#### 3.4. Provenance for a Lexical Entry and its Senses

The W3C Provenance Working Group defines provenance [49]:

*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 an 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 [1] (p. 6), [75]. 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 [1] (p. 6), [49,56,75,76].

The metadata used to describe a lexical entry is as follows [1] (p. 6–7):

- 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 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 *dct: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*.
- 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 U1 of the lexical entry *xh-n-isisu* is as follows:

```
: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  "xh-n-isisu"^^xsd:string ;
  rdfs:label     "isisu"@xh ;
  ontolex:canonicalForm :entry/xh-n-isisu#lemma ;
  ontolex:sense   :entry/xh-n-isisu#sense1 , :entry/xh-n-isisu#sense2 ;
  ontolex:denotes dbr:Abdomen , dbr:Stomach ;
  ontolex:evokes  :concept/0000000001 , :concept/0000000002 ;
  dct:isPartOf   :lexicon/xh ;
  prov:hadPrimarySource "The English-Xhosa Dictionary for Nurses"@en ;
  dct:license     <http://creativecommons.org/publicdomain/mark/1.0/> ;
  dct:creator     <https://londisizwe.org> ;
  prov:generatedAtTime "2017-09-19T05:00:00Z|+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:isPartOf :entry/xh-n-isisu ;
  prov:generatedAtTime "2017-09-19T05:00:00Z|+02:00"^^xsd:dateTime ;
  owl:versionInfo "2018-01-10"^^xsd:string ;
  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/000000002 ;
  dct:isPartOf :entry/xh-n-isisu ;
  prov:generatedAtTime "2018-01-10T05:00:00Z|+02:00"^^xsd:dateTime ;
  owl:versionInfo "2018-01-10"^^xsd:string ;
  owl:hasVersion :entry/xh-n-isisu/2018-01-10#sense2 .

```

In the lexical entry above, *prov:generatedAtTime* remains unchanged from the first recorded date and time for **U1**. However in **U3**, *prov:generatedAtTime* reflects the date and time that particular version was generated. Because of the *owl:sameAs* relation between **U1** and the latest version of **U3**, in order to account for the differing *prov:generatedAtTime* values for each, *prov:specializationOf* is used in **U3** to indicate that it “shares all aspects of” **U1**, and “additionally presents more specific aspects of the same thing” as **U1**, where an aspect in this context refers to **U3** as a version of **U1** generated at a particular point in time [49].

For the second version of **U3** onwards, *prov:wasRevisionOf* is used to indicate that that version was based on the previous version of **U3**, and within the same version, the previous version is indicated as outdated using *prov:invalidatedAtTime*. The generated RDF files for the lexical entry *xh-n-isisu* are dereferenceable at the following URIs:

**U1:** <https://londisizwe.org/entry/xh-n-isisu>  
**U3: Version 1:** <https://londisizwe.org/entry/xh-n-isisu/2017-09-19>  
**U3: Version 2:** <https://londisizwe.org/entry/xh-n-isisu/2018-01-10>

### 3.5. Modelling Provenance for a Lexicon

Using the same principles from the previous section, as well as the *lime* module from Ontolex-Lemon, the metadata used to describe a lexicon is as follows [1] (p. 7):

- Each lexicon is identified as a *lime:lexicon* and a *prov:Entity*.
- The *prov:generatedAtTime* property is recorded for each.
- The date a lexicon is changed is recorded using *dct:modified*.
- Other lexicons within the same namespace are indicated using *dct:references*.
- *owl:sameAs* is used to indicate that **U1** is the same as the latest version of **U3**.
- The version is indicated using *owl:versionInfo*.

- *dct:hasVersion* is used to show the previously generated versions, using the versioned URIs (U3 for lexicons).

The metadata only serves to describe the lexicon, and when a lexical entry is inserted or removed from a lexicon is not described [1] (p. 7). However, PROV-Dictionary, 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 [1] (p. 7), [77].

The generated RDF for version two of the lexicon *xh* follows below:

```
:lexicon/xh/2017-09-19/1
  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> ;
  lime:lexicalEntries"1"^^xsd:integer ;
  lime:linguisticCatalog <http://www.lexinfo.net/ontologies/2.0/lexinfo> ;
  dct:description"Londisizwe.org - isiXhosa lexicon"@en ;
  dct:license     <http://creativecommons.org/publicdomain/mark/1.0/> ;
  dct:creator     <https://londisizwe.org> ;
  prov:generatedAtTime    "2017-09-19T05:00:11Z|+02:00"^^xsd:dateTime ;
  prov:specializationOf   :lexicon/xh ;
  dct:modified           "2017-09-19"^^xsd:date ;
  owl:versionInfo"2017-09-19/1"^^xsd:string ;
  owl:hasVersion :lexicon/xh/2017-09-01/1 , :lexicon/xh/2017-09-19/1 ;
  dct:references :lexicon/en ;
  prov:derivedByInsertionFrom :lexicon/xh/2017-09-01/1 ;
  prov:qualifiedInsertion [
    a prov:Insertion;
    prov:dictionary :lexicon/xh/2017-09-01/1;
    prov:insertedKeyEntityPair [
      a prov:KeyEntityPair ;
      prov:pairKey "xh-n-isisu"^^xsd:string ;
      prov:pairEntity :entry/xh-n-isisu ;
    ] ;
  ] ;
] .

:lexicon/xh/2017-09-01/1
  prov:invalidatedAtTime "2017-09-19T05:00:11Z|+02:00"^^xsd:dateTime .
```

where:

- The current version was derived from the previous version, *:lexicon/xh/2017-09-01/1*, by means of inserting a key-value pair.
- The key, indicated above with the *prov:pairKey* relation, shares the same string literal as that for the *dct:identifier* relation in the associated lexical entry.
- The previous version of the lexicon is indicated to be outdated with the *prov:invalidatedAtTime* relation.
- Any information about the previous version beyond identifying it as a *prov:Dictionary* is not included here. Instead, that information will have been listed in the file of the previously published URI: <https://londisizwe.org/lexicon/xh/2017-09-01/1>

The generated RDF files for the lexicon *xh* are dereferenceable at the following URIs:

U1: <https://londisizwe.org/lexicon/xh>  
 U3: Version 1: <https://londisizwe.org/lexicon/xh/2017-09-01/1>  
 U3: Version 2: <https://londisizwe.org/lexicon/xh/2017-09-19/1>

Although McCrae et al. talk of grouping words in a lexicon as “no longer core” due to the linked data being published together on the web [40], should an external resource use the data beyond linking to it, it will be of value for the number of lexical entries in the lexicon to be identifiable. Due to the way PROV-Dictionary records insertions and removals, a published version cannot contain a mixture of insertions and removals, instead insertions should be recorded separately to removals, with the result that each published version of the lexicon will show a differing number of lexical entries to the previous version.

The class *prov:Dictionary* is defined as “an entity that provides a structure to some constituents, which are themselves entities [77]. 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” [77]. 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 [1] (p. 8). While this has not yet been explored for the EXDN dataset, it is work that could be considered in the future.

#### Step 10: Publish the RDF data

Content negotiation enables different representations of a resource to be presented, depending on the HTTP request of the client (web browser or software agent) [12] (pp. 26–27). Using content negotiation for EXDN, a human- and machine-readable view for each unversioned URI for lexical entries is presented; for versioned URIs of lexical entries and all URIs of lexicons and lexical concepts, the human-readable view defaults to the machine-readable view (RDF, serialised in Turtle), as these are more abstract concepts that are not expected to be relevant to the prototypical dictionary user.

For web servers running on Apache HTTP Server (Apache), the *.htaccess* file is used to configure content negotiation. An excerpt from the *.htaccess* file (for Apache Version 2.4) is shown below to demonstrate the implementation of content negotiation for a lexical concept:

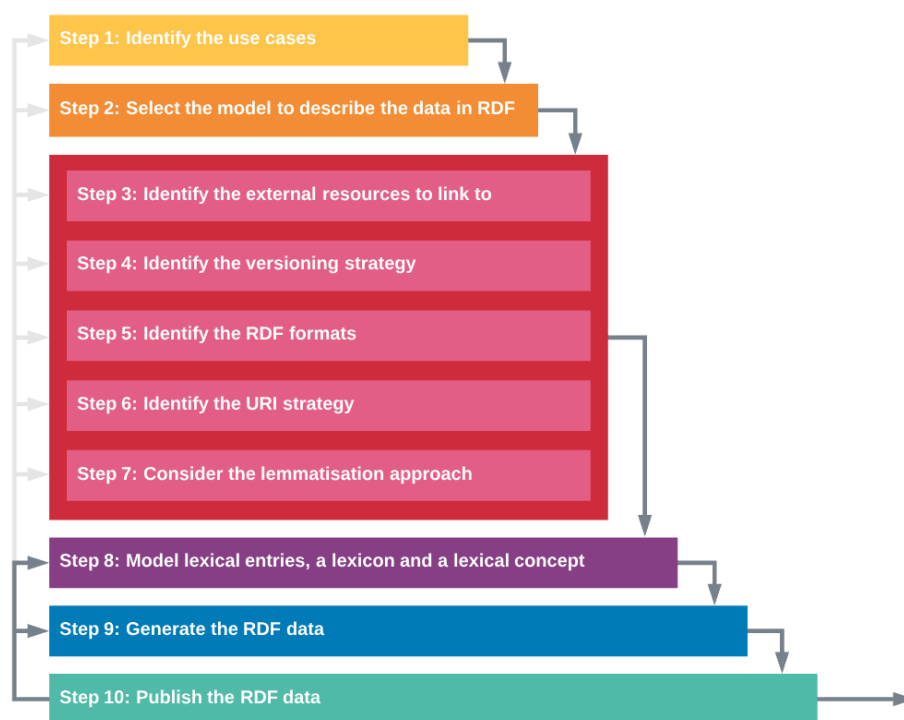
```
1 DirectoryCheckHandler On
2 RewriteEngine On
3 RewriteRule ^(.*)/$ /$1 [R,L]
4
5 RewriteCond %{HTTP_ACCEPT} text/turtle
6 RewriteRule ^concept/(.*)$ /rdf/concept/$1 [NC,R=302,L]
7 RewriteRule ^concept/(.*)$ /page/concept/$1 [NC,R=302,L]
8
9 RewriteCond %{REQUEST_FILENAME} !-f
10 RewriteRule ^rdf/(.*)$ /rdf/$1.ttl [L]
11 RewriteRule ^page/(.*)$ /script.php?file=$1 [L]
```

where:

- Line 1: By default, *DirectoryCheckHandler* is set to *Off*. Because the resource identifier for <https://londisizwe.org/concept/000000001> is the same as the directory name in <https://londisizwe.org/concept/000000001/2017-09-19>, Apache automatically appends a trailing slash to the resource identifier in the former URI; setting to *On* prevents this [78].
- Line 3: This removes a trailing slash appended to any URI/URLs.

- Line 5: This is a rule condition which checks if the HTTP header of the web browser or software agent is set to *text/turtle* [79]. If the condition is met, then the server proceeds to Line 6, otherwise the server proceeds to Line 7.
- Line 6: This is the RewriteRule which redirects the software agent to an RDF document of the original URI (thereby providing a machine-readable view of the URI), where the *[R]* flag indicates that an HTTP redirect is required [79,80].
- Line 7: This is the RewriteRule which redirects the web browser (or software agent which did not have an HTTP\_header of *text/turtle*) to a web page of the original URI (thereby providing a human-readable view of the URI) [79].
- Line 9: This is a rule condition which checks if the filename does not exist. For the URI <https://londisizwe.org/concept/000000001>, the filename would be 000000001. If the condition is met, the server proceeds to Line 10, otherwise it proceeds to Line 11 [79].
- Line 10: This is the RewriteRule which internally rewrites the URI for the RDF document from Line 6 to the URL as specified [81]. The URI from Line 6 will not change for the software agent.
- Line 11: This is the RewriteRule which internally rewrites the URI for the web page from Line 7 to the URL specified [81]. The URI from Line 7 will not change for the web browser / software agent. For selected browsers (such as Mozilla Firefox), when viewing a TTL file, despite a flag being set which changes the MIME type (for example, [T = text/plain,L]), an automatic download starts instead of it being displayed in the web browser [80]. To counter this, the URI is mapped to a script which retrieves the contents of the file and displays it to the end-user, with the Content-Type set to “text/plain” within the script [82].

To conclude this detailed discussion of each step, these methodological guidelines are expected to be iterative, and while every step could be revisited if necessary, it is expected that Steps 8–10 would account for the majority of change. The steps are also not strictly sequential, as shown in Figure 7, Steps 3–7 can be conducted in any order.



**Figure 7.** Conversion of a lexicographic resource to a Linguistic Linked Data Framework.



#### 4. Future Work and Conclusions

In the words of McArthur, a “printed and bound dictionary, ... is like a fossil; the moment it is complete and published, it is dated and rendered imperfect by the continuing flow of the language beyond what it has described.” [83] (p. 11). Using the methodological guidelines described in the previous section, the lexicographic resource, EXDN, has been taken from a state that is bounded and static, to a state that is unbounded and evolving. While its former state requires a human to infer meaning, the latter state, represented in RDF and described using Ontolex-Lemon, becomes machine-interoperable, offering the possibility for a machine to infer meaning. If one uses Ontolex-Lemon only so far as describing the lexicographic data, when converting from print form, separation between the lexical and the ontological layer (Requirement 2 in Step 2) is not strictly necessary as semantic representation can be loosely provided for using an external resource such as DBpedia, with more precise definitions given using string literals. However, if one wants to realise the purpose of Ontolex-Lemon, which is “to support [the] linguistic grounding of a given ontology” [28] with a view to multilingualism, then semantic representation would need to be defined using more formal ontologies. This leads to a discussion of future work.

Currently, the project is accessible via RDF crawling, as Linked Data documents, however, there is not a SPARQL endpoint. A cloud-hosted RDF platform will be used for the SPARQL endpoint, using a periodically updated RDF data dump, concatenated from the unversioned URIs of the Linked Data documents and each of its versions, serialised as N-Triples. To support this, alternative RDF formats will be generated, starting with N-Triples and JSON-LD. Longer-term, the intention is to evaluate Triple Pattern Fragments, a low-cost knowledge graph interface for live querying, proposed by Verborgh et al. [84].

Converting an isiXhosa lexical entry from an *ontolex:Word* to a *mmoon:DerivedNoun* is a time-consuming process, requiring identification of the prefixes, the plural form, and the noun classes for both the singular and plural forms. The use of crowdsourcing to assist with this work, integrated with a reputation management model, is planned for the latter part of 2019; crowdsourcing is also intended to be used to assist with determining the accuracy of the definitions from EXDN, as well as any definitions which have been machine translated from English using external resources in the public domain and imported into the project.

Versioning of each Linked Data document, as well as capturing its associated metadata, has been the primary focus of the LLDF proposed in this paper. As each document is text-based, comparison between the different versions of a lexical entry, for example, can be performed by simply comparing two versions, and indicating the list of lines that differ in both versions [85] (p. 8). The list of all versions for a resource is recorded in each unversioned URI, with the date included in each URI, however, this is insufficient as a version log. Future work will include a version log in the generated document of each unversioned URI, recording both the date and time for each version [86] (pp. 584–585). However, while this may indicate each version, what is not indicated is *what* has changed [86] (p. 584). For lexicons, by using *prov:qualifiedInsertion* and *prov:qualifiedDeletion* from PROV-Dictionary, the change in members is indicated from one version to the next, but for lexical entries and lexical concepts, a changelog will need to be implemented, with each change indicated to be an addition or a removal, accompanied by the affected triple [83] (p. 4).

To conclude, Ontolex-Lemon is presented as the “de facto standard” for modelling lexicographic resources [7] (p. 1), and as a model which is actively maintained and supported by comprehensive documentation, it has been found to be suitable for describing EXDN’s data in RDF, applying Linked Data principles, although the *semantics by reference* principle on which the model is based, is not without its drawbacks. EXDN is a dictionary of medical terms, where the meaning of terms can be represented using concepts from MeSH. However, ontology entities are not always available; an example is the lexical entries *Breath* and *Breathing* (both nouns) where the concept of a single breath is not available in MeSH, and if they had to be modelled using DBpedia, both would share the same denotation, namely *dbr:Breathing*.

Modelling multilinguality in RDF has been a challenging aspect. According to Fang et al. [38] (p. 51):

*translation relations can be inferred between terms in different languages when they refer to the same ontology entity. These lexical senses with an equivalent ontology reference have been regarded as a translation pair to be modelled.*

Indeed, an example of this was demonstrated in the introduction of this paper for the lexical entries *Abdomen* and *Isisu*. However, for the representation of more nuanced equivalence between the senses of different languages, an ontology entity is not sufficient, and context and contextualisations, identified by Gouws and Prinsloo to be of extreme importance in a bilingual dictionary when presenting translation equivalents, should be included [19] (p. 153).

Using Ontolex-Lemon's *vartrans* module, translation relations can be declared between a language pair, allowing for context and the use of different ontology entities for the source and target language [28], however, when a third language is introduced, there is the perception of redundancy when modelling a triple for each unidirectional translation relation. It is possible that the use of a relational database as the data store is contributing to this perception, and if the dictionary had to be Linked Data-native, as proposed by Garcia, Kernerman and Bosque-Gil, this perception may change [87].

As an alternative to modelling each unidirectional translation relation, the lexical concept was modelled as a shared conceptualisation for a sense from one or more languages; an example of which was demonstrated in Step 8. If one had to revisit Fang et al.'s statement, replacing "ontology entity" with "lexical concept", the statement would be more accurate.

As context and contextualisation is able to be modelled within the lexical concept, the result is that a concept is quite fine-grained, to the exclusion of near-synonyms. As mentioned in Step 8, a lexical concept is expressed by a member of a synset [72] (p. 665), and by including a reference to the member of an applicable synset, it allows more coarsely-grained near-synonyms to be associated with the lexical concept, thereby "grounding" the lexical concept. The implication of this though, is that where a synset does not exist for a given language, this would have to be created, with the lemma of the lexical entry serving as a singular member of a newly-created synset for that language. Within the context of EXDN, this has been done, although the resultant URI only serves to identify and is not dereferenceable.

For the ABD, there are 22 datasets, with each dataset containing translation relations between a language pair, for example English-Spanish [18] (p. 7). In order to construct a bilingual dictionary for which translations do not exist for a language pair in ABD, for example English-Portuguese, Gracia et al. talk of using an intermediate language to infer indirect translations, such as Spanish, where a dataset also exists for Spanish-Portuguese, and then calculating the confidence degree of the indirect translations from English-Portuguese [18] (p. 7). However, the construction of a dataset containing inferred language pairs can be similarly achieved using lexical concepts which contain multiple senses lexicalised to a concept, where each *ontolex:LexicalizedSense* can be derived from the target and source of a language pair or from lexical entries from monolingual dictionaries in the same domain.

By including a reference to a synset in a lexical concept, near-synonyms can be discovered for the languages of the senses lexicalised in the concept, however, the challenge remains that cross-lingual near-synonyms for languages of senses not lexicalised in the concept cannot be discovered. A tentative solution is to reference the Collaborative Interlingual Index (CILI), an index of WordNets, within the lexical concept by including an Interlingual Index (ILI) identifier, where an example ILI identifier is *ili:i66121* for the PWN synset {abdomen, stomach, belly, venter} [40] (pp. 591–593), [88]. This URI would take the human user or software agent to an external resource, however, if there is a SPARQL endpoint, lexical concepts referencing the same ILI identifier could be discovered within the resource, which in turn would enable other lexical entries to be discovered.

In closing, the methodological guidelines for converting a bilingual dictionary to an LLDF, with a view to multilinguality, have been described, where versioning and provenance has been the primary focus. Although the Ontolex-Lemon model has been integral to describing the lexical resource in RDF in a principled way, due to the versioning strategy, as well as the URI strategy which has deliberately remained agnostic of the model, should the model change, resulting in a new iteration from Step 8, or should the model *be* changed, resulting in a new iteration from Step 2, this change can be supported by the LLDF.

**Funding:** This research received no external funding.

**Acknowledgments:** I would like to thank my supervisor, Richard Higgs, for his commitment and support in seeing this research to its conclusion. I would also like to thank the anonymous reviewers for their kind and helpful feedback.

**Conflicts of Interest:** The author declares no conflict of interest.

## Appendix A

The following namespaces, shown here in Turtle RDF syntax, have been used in the code examples:

```
@prefix :           <https://londisizwe.org/> .
@prefix ontolox:    <http://www.w3.org/ns/lemon/ontolox#> .
@prefix lime:       <http://www.w3.org/ns/lemon/lime#> .
@prefix dbr:        <http://dbpedia.org/resource/> .
@prefix dct:        <http://purl.org/dc/terms/> .
@prefix foaf:       <http://xmlns.com/foaf/0.1/> .
@prefix lexinfo:    <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@prefix lonvoc:     <https://ontology.londisizwe.org/nounclass#> .
@prefix mesh:       <http://id.nlm.nih.gov/mesh/> .
@prefix mmoon:      <http://mmoon.org/core/> .
@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#> .
```

## References

1. Gillis-Webber, F. Managing provenance and versioning for an (evolving) dictionary in linked data format. In Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, Co-Located with LREC2018, Miyazaki, Japan, 7–12 May 2018. Available online: [http://lrec-conf.org/workshops/lrec2018/W23/pdf/2\\_W23.pdf](http://lrec-conf.org/workshops/lrec2018/W23/pdf/2_W23.pdf) (accessed on 10 October 2018).
2. Doke, C.M. *The Southern Bantu Languages*; International African Institute: London, UK, 1954.
3. Subfamily: Nguni (S.40). Available online: <http://glottolog.org/resource/languoid/id/ngun1276> (accessed on 11 February 2018).
4. Herbert, R.K.; Bailey, R. The Bantu languages: sociohistorical perspectives. In *Language in South Africa*; Mesthrie, R., Ed.; Cambridge University Press: Cambridge, UK, 2002; pp. 50–78.
5. Pretorius, L. The multilingual semantic web as virtual knowledge commons: the case of the under-resourced South African languages. In *Towards the Multilingual Semantic Web*; Buitelaar, P., Cimiano, P., Eds.; Springer: Berlin, Germany, 2014; pp. 49–66.
6. Taljard, E.; Bosch, S.E. A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written Bantu languages. *Nord. J. Afr. Stud.* **2006**, *15*, 428–442.

7. Bosque-Gil, J.; Gracia, J.; Montiel-Ponsoda, E. Towards a module for lexicography in OntoLex. In Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017), Galway, Ireland, 18 June 2017. Available online: [http://ceur-ws.org/Vol-1899/OntoLex\\_2017\\_paper\\_5.pdf](http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_5.pdf) (accessed on 20 October 2018).
8. Crystal, D. *The Cambridge Encyclopedia of Language*; Cambridge University Press: Cambridge, UK, 2010.
9. Cyganiak, R.; Wood, D.; Lanthaler, M. *RDF 1.1 Concepts and Abstract Syntax—W3C Recommendation 25 February 2014*; World Wide Web Consortium. Available online: <https://www.w3.org/TR/rdf11-concepts/> (accessed on 15 October 2018).
10. Tim Berners-Lee: The Next Web. 2009. Available online: [http://www.ted.com/talks/tim\\_berniers\\_lee\\_on\\_the\\_next\\_web.html](http://www.ted.com/talks/tim_berniers_lee_on_the_next_web.html) (accessed on 15 April 2017).
11. Berners-Lee, T. Linked Data. 2006. Available online: <https://www.w3.org/DesignIssues/LinkedData.html> (accessed on 25 December 2017).
12. Hyvönen, E. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012.
13. Van Hooland, S.; Verborgh, R. *Linked Data for Libraries, Archives and Museums*; Facet Publishing: London, UK, 2014.
14. Wood, D.; Zaidman, M.; Ruth, L.; Hausenblas, M. *Linked Data: Structured Data on the Web*; Manning Publications Co: New York, NY, USA, 2014.
15. Gracia, J. Introduction to linked data for language resources. In Proceedings of the 2nd Summer Datathon on Linguistic Linked Open Data, Cercedilla, Spain, 26–30 June 2017.
16. About | DBpedia. Available online: <https://wiki.dbpedia.org/about> (accessed on 10 January 2018).
17. Converting BabelNet as Linguistic Linked Data. Available online: [https://www.w3.org/community/bpmlod/wiki/Converting\\_BabelNet\\_as\\_Linguistic\\_Linked\\_Data](https://www.w3.org/community/bpmlod/wiki/Converting_BabelNet_as_Linguistic_Linked_Data) (accessed on 5 December 2017).
18. Gracia, J.; Villegas, M.; Gómez-Pérez, A.; Bel, N. The Apertium bilingual dictionaries on the web of data. *Semant. Web* **2018**, *9*, 231–240. Available online: <http://www.semantic-web-journal.net/system/files/swj1419.pdf> (accessed on 31 December 2017). [CrossRef]
19. Gouws, R.H.; Prinsloo, D.J. *Principles and Practice of South African Lexicography*; SUN MeDIA: Stellenbosch, South Africa, 2005.
20. Grace's Guide to British Industrial History: Bengers Food. Available online: [https://www.gracesguide.co.uk/Bengers\\_Food](https://www.gracesguide.co.uk/Bengers_Food) (accessed on 12 October 2018).
21. Haushofer, L. Between food and medicine: artificial digestion, sickness, and the case of Benger's Food. *J. Hist. Med. Allied Sci.* **2018**, *73*, 168–187. [CrossRef] [PubMed]
22. Villazón-Terrazas, B.; Vilches-Blázquez, L.M.; Corcho, O.; Gómez-Pérez, A. Methodological guidelines for publishing government linked data. In *Linking Government Data*; Wood, D., Ed.; Springer: New York, NY, USA, 2012; pp. 27–49. Available online: [https://link.springer.com/chapter/10.1007/978-1-4614-1767-5\\_2](https://link.springer.com/chapter/10.1007/978-1-4614-1767-5_2) (accessed on 12 October 2018).
23. Vila-Suero, D.; Gómez-Pérez, A.; Montiel-Ponsoda, E.; Gracia, J.; Aguado-de-Cea, G. Publishing linked data on the web: The multilingual dimension. In *Towards the Multilingual Semantic Web*; Buitelaar, P., Cimiano, P., Eds.; Springer: Berlin, Germany, 2014; pp. 101–117.
24. Gracia, J.; Vila-Suero, D. *Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries*; Final Community Group Report 29 September 2015; W3C Best Practices for Multilingual Linked Open Data Community Group under the W3C Community Final Specification Agreement (FSA), World Wide Web Consortium: 2015. Available online: <https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/> (accessed on 25 December 2017).
25. Zainal, Z. Case study as a research method. *Jurnal Kemanusiaan* **2007**, *9*, 1–6.
26. 2nd Summer Datathon on Linguistic Linked Open Data (SD-LLOD-17). Available online: <http://datathon2017.retele.linkeddata.es/> (accessed on 12 October 2018).
27. Lemon—The Lexicon Model for Ontologies. Available online: <https://lemon-model.net/> (accessed on 10 September 2018).
28. Lexicon Model for Ontologies: Community Report, 10 May 2016. Final Community Group Report 10 May 2016, W3C Ontology-Lexica Community Group under the W3C Community Final Specification Agreement (FSA), World Wide Web Consortium: 2016. Available online: <https://www.w3.org/2016/05/ontolex/> (accessed on 19 December 2017).

29. Ontology-Lexica Community Group. Available online: <https://www.w3.org/community/ontolex/> (accessed on 12 October 2018).
30. McCrae, J.P.; Unger, C. Design patterns for engineering the ontology-lexicon interface. In *Towards the Multilingual Semantic Web*; Buitelaar, P., Cimiano, P., Eds.; Springer: Berlin, Germany, 2014; pp. 15–30.
31. Francopoulo, G.; George, M. Model description. In *LMF—Lexical Markup Framework*; Francopoulo, G., Ed.; ISTE Ltd.: London, UK, 2013.
32. McCrae, J. LMF. 2012. Available online: <http://lemon-model.net/lemon-cookbook/node46.html> (accessed on 20 October 2018).
33. McCrae, J.; Spohr, D.; Cimiano, P. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*; Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 245–259.
34. Faab, G.; Bosch, S.E.; Gouws, R.H. A general lexicographic model for a typological variety of dictionaries in African languages. *Lexikos* **2014**, *24*, 94–115.
35. Cimiano, P.; Buitelaar, P.; McCrae, J.; Sintek, M. LexInfo: A declarative model for the lexicon-ontology interface. *Web Semant. Sci. Serv. Agents World Wide Web* **2010**, *9*, 29–51. [[CrossRef](#)]
36. Montiel-Ponsoda, E.; Vila-Suero, D.; Villazón-Terrazas, B.; Dunsire, G.; Escolano Rodríguez, E.; Gómez-Pérez, A. Style guidelines for naming and labeling ontologies in the multilingual web. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2011, The Hague, The Netherlands, 21–23 September 2011*. Available online: [http://oa.upm.es/12469/1/INVE\\_MEM\\_2011\\_105132.pdf](http://oa.upm.es/12469/1/INVE_MEM_2011_105132.pdf) (accessed on 26 October 2018).
37. Espinoza, M.; Gómez-Pérez, A.; Montiel-Ponsoda, E. Multilingual and localization support for ontologies. In *The Semantic Web: Research and Applications*; Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 821–825.
38. Fang, Z.; Wang, H.; Gracia, J.; Bosque-Gil, J.; Ruan, T. Zhishi.lemon: On publishing Zhishi.me as linguistic linked open data. In *The Semantic Web: ISWC 2016*; Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y., Eds.; Springer: Cham, Switzerland, 2016; pp. 47–55.
39. Khalfi, M.; Nahli, O.; Zarghili, A. Classical dictionary Al-Qamus in lemon. In *Proceedings of the 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, Tangier, Morocco, 24–26 October 2016. [[CrossRef](#)]
40. McCrae, J.P.; Bosque-Gil, J.; Gracia, J.; Buitelaar, P.; Cimiano, P. The Ontolex-Lemon model: Development and applications. In *Proceedings of the eLex 2017 Electronic Lexicography in the 21st Century: Lexicography from Scratch*, Leiden, The Netherlands, 19–21 September 2017. Available online: <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf> (accessed on 26 October 2018).
41. Tittel, S.; Chiarcos, C. Historical lexicography of Old French and linked open data: Transforming the resources of the Dictionnaire étymologique de l’ancien français with Ontolex-Lemon. In *Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, Co-Located with LREC2018*, Miyazaki, Japan, 12 May 2018. Available online: [http://lrec-conf.org/workshops/lrec2018/W23/pdf/2\\_W33.pdf](http://lrec-conf.org/workshops/lrec2018/W23/pdf/2_W33.pdf) (accessed on 15 October 2018).
42. DCMI Metadata Terms. Dublin Core Metadata Initiative: 2012. Available online: <http://dublincore.org/documents/dcmi-terms/> (accessed on 10 January 2018).
43. Brickley, D.; Miller, L. FOAF Vocabulary Specification 0.99. 2014. Available online: <http://xmlns.com/foaf/spec/> (accessed on 10 January 2018).
44. Library of Congress Names. Available online: <http://id.loc.gov/authorities/names.html> (accessed on 10 January 2018).
45. Library of Congress Subject Headings. Available online: <http://id.loc.gov/authorities/subjects.html> (accessed on 10 January 2018).
46. Wunner, T. LEXINFO Vocabulary. DERI Vocabularies: 2012. Available online: <http://vocab.deri.ie/lexinfo#> (accessed on 17 January 2018).
47. Fact Sheet: Medical Subject Headings. Available online: <https://www.nlm.nih.gov/pubs/factsheets/mesh.html> (accessed on 10 January 2018).



48. The Multilingual Morpheme Ontology: Home. Available online: <http://mmoon.org/> (accessed on 17 January 2018).
49. Lebo, T.; Sahoo, S.; McGuinness, D.; Belhajjame, K.; Cheney, J.; Corsar, D.; Garijo, D.; Soiland-Reyes, S.; Zednik, S.; Zhao, J. PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013, World Wide Web Consortium: 2013. Available online: <https://www.w3.org/TR/prov-o/> (accessed on 1 January 2018).
50. WordNet RDF. Available online: <http://wordnet-rdf.princeton.edu/> (accessed on 11 November 2017).
51. Alexander, K.; Cyganiak, R.; Hausenblas, M.; Zhao, J. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note 03 March 2011, World Wide Web Consortium: 2011. Available online: <https://www.w3.org/TR/void/> (accessed on 10 January 2018).
52. Di Maio, P. Linked data beyond libraries. In *Linked Data and User Interaction*; Cervone, H.F., Svensson, L.G., Eds.; Walter de Gruyter GmbH: Berlin, Germany, 2015.
53. McCrae, J.; Montiel-Ponsoda, E.; Cimiano, P. Integrating WordNet and Wiktionary with lemon. In *Linked Data in Linguistics*; Chiarcos, C., Nordhoff, S., Hellman, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 25–34.
54. Bouda, P.; Cysouw, M. Treating dictionaries as a linked-data corpus. In *Linked Data in Linguistics*; Chiarcos, C., Nordhoff, S., Hellman, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 15–24.
55. Klein, M.; Fensel, D. Ontology versioning on the Semantic Web. In *Proceedings of the First International Conference on Semantic Web Working*, California, CA, USA, 30 July–1 August 2001. Available online: <https://pdfs.semanticscholar.org/417f/b1dd895a9416f9d56932e6b3870749ba582c.pdf> (accessed on 18 October 2018).
56. Flati, T.; Moro, A.; Matteis, L.; Navigli, R.; Velardi, P. Guidelines for linguistic linked data generation: Multilingual dictionaries (Babelnet). Final Community Group Report 29 September 2015, W3C Best Practices for Multilingual Linked Open Data Community Group under the W3C Community Final Specification Agreement (FSA), World Wide Web Consortium: 2015. Available online: <https://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/> (accessed on 27 December 2017).
57. Van Erp, M. Reusing linguistic resources: Tasks and goals for a linked data approach. In *Linked Data in Linguistics*; Chiarcos, C., Nordhoff, S., Hellman, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 57–64.
58. Eckart, K.; Riester, A.; Schweitzer, K. A discourse information radio news database for linguistic analysis. In *Linked Data in Linguistics*; Chiarcos, C., Nordhoff, S., Hellman, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 65–76.
59. De Rooij, S.; Beek, W.; Bloem, P.; van Harmelen, F.; Schlobach, S. Are names meaningful? Quantifying social meaning on the semantic web. In *The Semantic Web: ISWC 2016*; Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 184–199.
60. Kiryakov, A.; Ognyanov, D. Tracking changes in RDF(S) repositories. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002)*, Sig enza, Spain, 1–4 October 2002; Gómez-Pérez, A., Benjamins, V.R., Eds.; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2473, pp. 373–378. Available online: <https://pdfs.semanticscholar.org/9dec/6fddad51df8d708dfe5b97b752fb563fc08a.pdf> (accessed on 20 October 2018).
61. Bond, F.; Vossen, P.; McCrae, J.P.; Fellbaum, C. CILI: The Collaborative Interlingual Index. 2016. Available online: <http://gwc2016.racai.ro/Slide-uri/day01/Bond,%20The%20Collaborative%20Interlingual%20Index.pdf> (accessed on 18 October 2018).
62. Heath, T.; Bizer, C. *Linked Data: Evolving the Web into a Global Data Space*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2011.
63. Archer, P.; Goedertier, S.; Loutas, N. D7.1.3—Study on Persistent URIs, with Identification of Best Practices and Recommendations on the Topic for the MSs and the EC. 2012. Available online: <https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf> (accessed on 26 December 2017).
64. Hogan, A.; Umbrich, J.; Harth, A.; Cyganiak, R.; Polleres, A.; Decker, S. An empirical survey of linked data conformance. *Web Semant. Sci. Serv. Agents World Wide Web* **2012**, *14*, 14–44. [CrossRef]
65. Simons, N.; Richardson, J. *New Content in Digital Repositories: The Changing Research Landscape*; Chandos Publishing: Oxford, UK, 2013.

66. Keller, M.A.; Persons, J.; Glaser, H.; Calter, M. Report on the Stanford Linked Data Workshop, 27 June–1 July 2011. Available online: <https://www.clir.org/wp-content/uploads/sites/6/LinkedDataWorkshop.pdf> (accessed on 26 December 2017).
67. Labra Gayo, J.E.; Kontokostas, D.; Auer, S. Multilingual Linked Data Patterns. *Semant. Web J.* **2015**, *6*. Available online: <http://www.semantic-web-journal.net/system/files/swj495.pdf> (accessed on 27 December 2017). [CrossRef]
68. Sachs, J.; Finin, T. What Does It Mean for a URI to resolve? In Proceedings of the AAAI Spring Symposium on Linked Data Meets Artificial Intelligence, Palo Alto, CA, USA, 2010. Available online: [http://ebiquity.umbc.edu/\\_file\\_directory\\_/papers/495.pdf](http://ebiquity.umbc.edu/_file_directory_/papers/495.pdf) (accessed on 26 December 2017).
69. Prinsloo, D. Review: Oxford Bilingual School Dictionary: Zulu and English. *Lexikos* **2010**, *20*, 760–766. [CrossRef]
70. De Schryver, G.-M. Revolutionizing Bantu lexicography—A Zulu case study. *Lexikos* **2010**, *20*, 161–201. [CrossRef]
71. Zgusta, L. *Manual of Lexicography*; Academia, Publishing House of the Czechoslovak Academy of Sciences: Prague, Czech Republic, 1971.
72. Cookbook for Open Government Linked Data. Available online: [https://www.w3.org/2011/gld/wiki/Linked\\_Data\\_Cookbook](https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook) (accessed on 4 January 2018).
73. MacVicar, N. “Breath”. In *English-Xhosa Dictionary for Nurses*, 2nd ed.; Lovedale Press: Lovedale, South Africa, 1935; p. 13.
74. McCrae, J.P.; Gracia, J. Introduction to the Ontolex-Lemon Model. In Proceedings of the 2nd Summer Datathon on Linguistic Linked Open Data, Cercedilla, Spain, 26–30 June 2017.
75. Faniel, I.M.; Yakel, E. 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*; Johnston, L.R., Ed.; Association of College and Research Libraries: Chicago, IL, USA, 2017; pp. 103–126.
76. Tennis, J.T. Scheme versioning in the semantic web. In *Knitting the Semantic Web*; Greenberg, J., Méndez, E., Eds.; CRC Press: Boca Raton, FL, USA, 2007; pp. 85–104.
77. Missier, P.; Moreau, L.; Cheney, J.; Lebo, T.; Soiland-Reyes, S. PROV-Dictionary: Modeling Provenance for Dictionary Data Structures. W3C Working Group Note 30 April 2013, World Wide Web Consortium: 2013. Available online: <https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/> (accessed on 1 January 2018).
78. Apache Module Mod\_Dir. Available online: [https://httpd.apache.org/docs/2.4/mod/mod\\_dir.html](https://httpd.apache.org/docs/2.4/mod/mod_dir.html) (accessed on 20 October 2018).
79. Apache Module Mod\_Rewrite. Available online: [https://httpd.apache.org/docs/2.4/mod/mod\\_rewrite.html](https://httpd.apache.org/docs/2.4/mod/mod_rewrite.html) (accessed on 20 October 2018).
80. RewriteRule Flags. Available online: <https://httpd.apache.org/docs/2.4/rewrite/flags.html> (accessed on 20 October 2018).
81. Redirecting and Remapping with Mod\_Rewrite. Available online: <https://httpd.apache.org/docs/2.4/rewrite/remapping.html> (accessed on 20 October 2018).
82. PHP: header—Manual. Available online: <http://php.net/manual/en/function.header.php> (accessed on 31 October 2018).
83. McArthur, T. *Worlds of Reference*; Cambridge University Press: Cambridge, UK, 1986.
84. Verborgh, R.; Vander Sande, M.; Hartig, O.; Van Herwegen, J.; De Vocht, L.; De Meester, B.; Haesendonck, G.; Colpaert, P. Triple pattern fragments: A low-cost knowledge graph interface for the web. *J. Web Semant.* **2016**, *37*, 184–206. [CrossRef]
85. Noy, N.F.; Musen, M.A. Ontology versioning in an ontology management framework. *IEEE Intell. Syst.* **2004**, *19*, 6–13. [CrossRef]
86. Plessers, P.; De Troyer, O. Ontology change detection using a version log. In Proceedings of the 4th International Conference on The Semantic Web, Galway, Ireland, 6–10 November 2005. Available online: <https://pdfs.semanticscholar.org/3c52/491aa37b6291b58630de25bcd8f2262aebb5.pdf> (accessed on 19 October 2018).



87. Gracia, J.; Kernerman, I.; Bosque-Gil, J. Toward linked data-native dictionaries. In Proceedings of the eLex 2017 Electronic Lexicography in the 21st Century: Lexicography from Scratch, Leiden, The Netherlands, 19–21 May 2017. Available online: <https://elex.link/elex2017/wp-content/uploads/2017/09/paper33.pdf> (accessed on 22 October 2018).
88. Princeton WordNet 3.1. Available online: <https://wordnet-rdf.princeton.edu/ttl/lemma/abdomen> (accessed on 24 October 2018).



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).