LIDER: FP7 – 610782

Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe

Deliverable number	**D2.1.1**
Deliverable title	**Guidelines and best practices for Linguistic Linked Data-based content analytics – Phase I**
Main Authors	**Philipp Cimiano, John McCrae, Jorge Gracia, Bettina Klimek, Martin Brümmer, Ciro Baroni, Dave Lewis, Victor Rodríguez Doncel, Roberto Navigli, Tiziano Flati**

Grant Agreement number	610782
Project ref. no	FP7-610782
Project acronym	LIDER
Project full name	Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe
Starting date (dur.)	1/11/2013 (24 months)
Ending date	31/10/2015
Project website	http://www.lider-project.eu/

Coordinator	Asunción Gómez-Pérez
Address	Campus de Montegancedo sn. 28660 Boadilla del Monte, Madrid, Spain
Reply to	asun@fi.upm.es
Phone	+34-91-3367417
Fax	+34-91-3524819

Document Identifier	D2.1.1
Class Deliverable	LIDER EU-ICT-2013-610782
Version	2.0
Document due date	September 2014
Submitted	October 2014
Responsible	Dr. Dave Lewis, Trinity College Dublin (TCD)
Reply to	Dave.Lewis@scss.tcd.ie
Document status	Final
Nature	R(Report)
Dissemination level	PU(Public)
WP/Task responsible(s)	WP2, Tasks 2.1 - 2.3
Contributors	Philipp Cimiano, John McCrae, Jorge Gracia, Bettina Klimek, Martin Brümmer, Ciro Baroni, Dave Lewis, Victor Rodríguez Doncel
Distribution List	Consortium Partners
Reviewers	Paul Buitelaar, NUIG
Document Location	http://lider-project.eu/?q=doc/deliverables

Executive Summary

In this deliverable we have presented guidelines for the publication of multilingual data as linked data. It includes guidelines on the appropriate use of existing vocabularies; the naming of resources; dereferencing resources; encoding textual content; interlinking resources and language identification. It also captures developing best practices garnered from mapping meta-data in existing language resource repositories into linked data. It provides detailed guidelines in mapping major classes of lexical resources and dictionaries into linked data, using the LEMON lexical-semantic vocabulary and the NLP Interchange Format (NIF) as a common base.

As a platform for the further development and application of best practice, a critical comparison of existing linguistic meta-data repositories is conducted so as to indicate a path for meta-data harmonisation between these major resources.

The work presented here is the result of widespread consultation and engagement with the relevant stakeholder communities. This engagement included the active gathering of requirements and use cases; direct engagement with the communities operating the existing linguistic resource meta-data repositories and ongoing opportunities for influencing the development of technical best practice and linked data vocabulary recommendation through W3C community groups active in this area. This document therefore provides just a snapshot of many ongoing activities and the reader is encouraged to engage with these directly through the links provided.

Document Information

IST Project Number	FP7-610782	Acronym	LIDER
Full Title	Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe		
Project URL	http://www.lider-project.eu/		
Document URL	http://lider-project.eu/?q=doc/deliverables		
EU Project Officer	Susan Fraser		

Deliverable	Number	D2.1.1	Title	Guidelines and best practices for Linguistic Linked Data-based content analytics – Phase I
Workpackage	Number	WP2	Title	Guidelines and best practices for industry

Date of Delivery	Contractual	1st October 2014	Actual	15th October 2014
Status	version 2		final ■	
Nature	prototype □ report ■ dissemination □			
Dissemination level	public ■ consortium □			

Authors (Partner)	Philipp Cimiano, John McCrae, Jorge Gracia, Bettina Klimek, Martin Brümmer, Ciro Baroni, Dave Lewis, Victor Rodríguez Doncel			
Responsible Author	Name	Kevin Koidl	E-mail	kevin.koidl@scss.tcd.ie
	Partner	TCD	Phone	-

Abstract (for dissemination)	In this deliverable we have presented guidelines for the publication of multilingual data as linked data. It includes guidelines on the appropriate use of existing vocabularies; the naming of resources; dereferencing resources; encoding textual content; interlinking resources and language identification. It also captures developing best practices garnered from mapping meta-data in existing language resource repositories into linked data. It provides detailed guidelines in mapping major classes of lexical resources and dictionaries into linked data, using the LEMON lexical-semantic vocabulary and the NLP Interchange Format (NIF) as a common base. As a platform for the further development and application of best

	practice, a critical comparison of existing linguistic meta-data repositories is conducted so as to indicate a path for meta-data harmonisation between these major resources. The work presented here is the result of widespread consultation and engagement with the relevant stakeholder communities. This engagement included the active gathering of requirements and use cases; direct engagement with the communities operating the existing linguistic resource meta-data repositories and ongoing opportunities for influencing the development of technical best practice and linked data vocabulary recommendation through W3C community groups active in this area. This document therefore provides just a snapshot of many ongoing activities and the reader is encouraged to engage with these directly through the links provided.
Keywords	Linguistic linked data, best practice, RDF, lexicons, meta-data

Version	Modification(s)	Date	Author(s)
01	First initial version for collaborative editing	15/09/2014	Philipp Cimiano, John McCrae, Jorge Gracia, Bettina Klimek, Martin Brümmer, Ciro Baroni, Dave Lewis, Victor Rodríguez Doncel
02	Final version	15/10/2014	Kevin Koidl

Project Consortium Information

Participants		Contact
Universidad Politécnica de Madrid		Asunción Gómez-Pérez Email: asun@fi.upm.es
The Provost, Fellows, Foundation Scholars & The Other Members of Board of The College of the Holy & Undivided Trinity of Queen Elizabeth near Dublin (Trinity College Dubl, Ireland)		David Lewis Email: dave.lewis@cs.tcd.ie
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI, Germany)		Felix Sasaki Email: fsasaki@w3.org
National University of Ireland, Galway (NUI Galway, Ireland)		Paul Buitelaar Email: paul.buitelaar@deri.org
Institut für Angewandte Informatik EV (INFAI, Germany)		Sebastian Hellmann Email: hellmann@informatik.uni-leipzig.de
Universität Bielefeld (UNIBI, Germany)		Philipp Cimiano Email: cimiano@cit-ec.uni-bielefeld.de
Universita degli Studi di Roma La Sapienza (UNIVERSITA DEGLI STU, Italy)		Roberto Navigli Email: navigli@di.uniroma1.it
GEIE ERCIM (ERCIM, France)		Felix Sasaki Email: fsasaki@w3.org

Table of Contents

1 Introduction

The goal of WP2 is to provide a set of guidelines and best practices that support the generation of Linguistic Linked Data (LLD) and its exploitation in content analytics. This deliverable describes guidelines that the LIDER project has developed along three levels:

- **Identification of best practices and vocabularies for metadata for multilingual and multimedia web content (Task 2.1):** This task is concerned with the identification of models and best practices that support the description of multilingual and multimedia content on the Web. In the first phase we have concentrated on *identifying relevant vocabularies that are recommended to be used for the description of linguistic linked data sets*. Aspects related to the description of multimedia content have been left out in this first phase. In particular, we have concentrated on the vocabularies used in the guidelines for *Linguistic Linked Data (LLD)* generation (see below). We also provide a set of general recommendations to be applied for modelling and generating multilingual linked data as developed by the Best Practices for Multilingual Linked Data (BPMLOD) community group. Finally, we also describe the work done in the context of defining a so called *DataID* as a uniform way to describe the general metadata of datasets.
- **Development of guidelines and models for Linguistic Linked Data generation, publication and exploitation (Task 2.2):** The goal of this task is the development of guidelines that support the entire lifecycle of linguistic linked data resources, starting from i) the **modelling and generation of linked data resources,** ii) the **publication** of these resources on the web, to iii) the **exploitation** of these resources in content analytic tasks. In this first phase, we have focused on the development for guidelines for the generation and publication of a number of frequent and relevant types of linguistic resources, including Wordnet, bilingual dictionaries, terminological resources (in TBX format), BabelNet, as well as natural language processing services. The latter is a first step towards developing content analytics services that exploit linguistic linked data available on the Web.
- **Development of guidelines for LLD-aware NLP services (Task 2.3):** The goal of this task is to develop guidelines to enable content analytics processes to discover LLD resources by means of querying the Web using search engines and data repositories. In the first phase of the project, to support discovery of relevant linguistic resources by LLD-aware NLP services, we have analyzed current repositories and attempted to homogenize the vocabularies used by different repositories of linguistic metadata as a basis to support the discovery of relevant resources.

The deliverable is structured as follows. Section 1 describes our work with respect to the identification of best practices and vocabularies for metadata for multilingual LLOD.

Section 2 describes the development of guidelines and models for Linguistic Linked Data (LLD) generation and publication (Task 2.2). Section 3 describes the development of guidelines for LLD-aware NLP services and analyzes in particular the state of play of current repositories of metadata for linguistic datasets (Task 2.3). Section 4 provides a short conclusion and describes next steps.

2 Requirements and Use Cases

The best practice guidelines activities identified above are scoped and guided from the requirements and use cases gather by WP1.

These requirements and use cases have been gathered through direct engagement with stakeholders through a W3C Community Group established to gather, disseminate and gather feedback on requirements and use cases. The community is titled Linked Data for Language Technology (LD4LT) and all of its activities are open to the public via its community portal[1] hosted by the W3C. The LD4LT community is supported by LIDER via the activities of WP4. It provides an online hub and remote conferencing opportunities for the requirements and use case related activities. It is also the public organisational base for publicising and disseminating the results from a series of face to face RoadMapping workshops and surveys. The group has also assembles reference to relevant use cases established by other projects and communities and has conducted a number of surveys to obtain a broader view of opinions. Finally, the LD4LT Community provide a technical consensus building forum for addressing immediate and strategically important interoperability issues that are related to the future development and deployment for linguistic linked data. The LD4LT Community also focuses on liaising with other groups and supporting them in developing and publishing linguistic linked data in line with best practice (presented in this document) and technical architecture. This provides further direct external validation of these outputs of the LIDER project.

2.1 Analysis of Requirements and Use Case

The requirements and use cases gathered through the LD4LT outreach and community engagement can be broadly characterised under three heading:

- **Global Customer Engagement Use Cases:** This reflects use cases offered by commercial organisations. These address different aspects of how companies interact with their customers with global markets across different linguistic and cultural norms. This involves the translation and localisation of content generated by companies for consumption by customers or potential customers and support for content search across those languages. This typically requires domain specific multilingual language resources to support language technology such as machine translation and search and indexing. Increasingly however this also involves the ability to analyse content generated by customers and other third parties as they comment on, review, pose questions about or provide answers on specific products and services via numerous digital channels and languages.

[1] http://www.w3.org/community/ld4lt/

Providers of specialised digital support services, such as language services and content analytics are important sources of use cases, reflecting the growth and innovation in value chains in bringing language resources and technology to commercial applications.

- **Public Sector and Civil Society Use Cases:** the Public Sector has been an early adopter of linked data, emphasising the use of linked *open* data motivated by transparency requirements and open data obligations that are increasingly common in national and transnational public administration. Such open data may include content which may benefit from linguistic annotation or which may serve as linguistic corpora, e.g. DG-T annual release of it translation memory, which is the most popular download from the European Commission's Open Data portal. The public sector, non-governmental organisations, non-profits representing specific domains, and academia also work to curate high quality language resource, including dictionaries and lexicons for public consumption, access to which may be enhanced via linguistic linked data techniques. Finally, large-scale communities organised as international non-profits are also providing major crowd-sourced language resources. While these bodies are also interested in adopting language technologies, their financial resources are limited so the emphasis is on the availability of open source solutions that are compatible with available language resources.
- **Linguistic Linked Data Life Cycle and Value Network Requirements:** While individual commercial, public sector and other civil society actors are typically focussed on their own use cases, common themes often emerge that highlight dependencies between organisations in publishing, discovering, using and enhancing linguistic linked data as an asset with value in content processing, content analytics and the application of language technology. These highlight the need for a life-cycle view of linguistic linked data. This can help explain how language data quality interacts with the value it provides to different actors and how the costs involved impact of the value, e.g. resource licensing, overcoming technical interoperability barriers, evaluating quality and compliance to data protection rules. These issues are important in planning how the development of linguistic linked data can support global markets in digital good and services. The EC in particular seeks to identify how digital good and services can be leveraged by companies, especially European SMEs, within EU internal market. Further the EC seeks to support the internal market with Digital Service under the Connecting Europe Facilities. These include several facilities where the processing and analytics of multilingual content and unstructured data will be key and therefore where efficient markets or pooling strategies for language resources are required.

To date the best practices are based on experiences with linguistic linked data from the public sector and civil society use cases, as they represent the more mature applications of linguistic linked data. Several of the best practice guidelines presented for generation, publication and exploitation are therefore tuned for use cases related to existing public domain data. However, the articulation of these best practices has been conducted with

a growing understanding of commercial use cases, including concerns about licensing and quality of meta-data. The general best practices presented here, therefore, are designed to be applicable also to more commercial use cases related to global customer engagement and their adoption are necessary (though not sufficient) for the development of international pooling facilities and return of investment decisions for linguistic linked data. Some of the generation, publication and exploitation guidelines are also focussed on commercial use cases, in particular those addressing Terminology in the TBX format and the use of multilingual dictionaries. Finally the harmonisation of existing repositories is the basis for both reducing the cost of accessing language resources as linguistic linked data and to understanding the best practices for seamless pooling and reuse as well as for future digital markets for language resources.

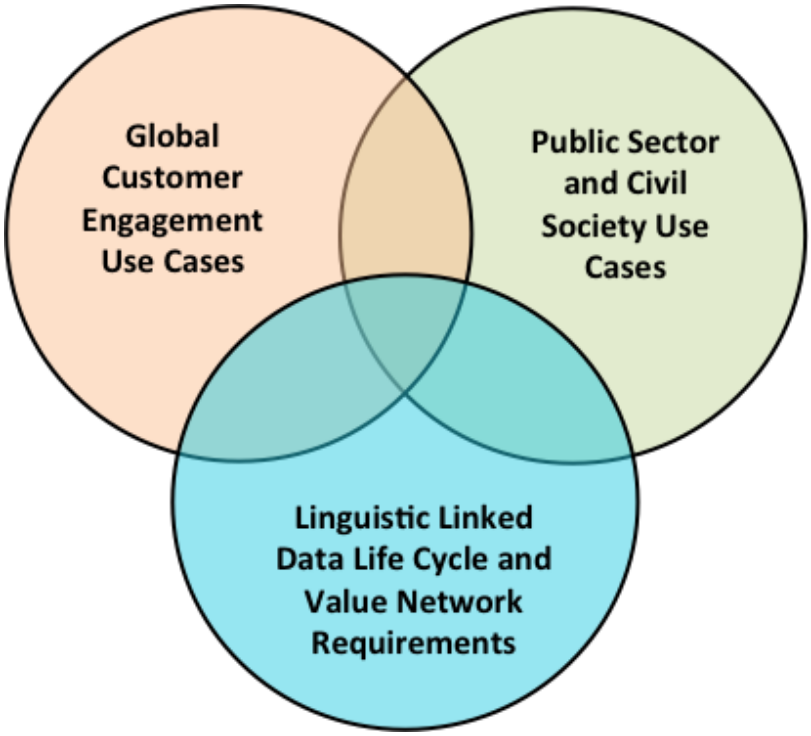


Figure 1 Structure of LD4LT requirements and use case analysis

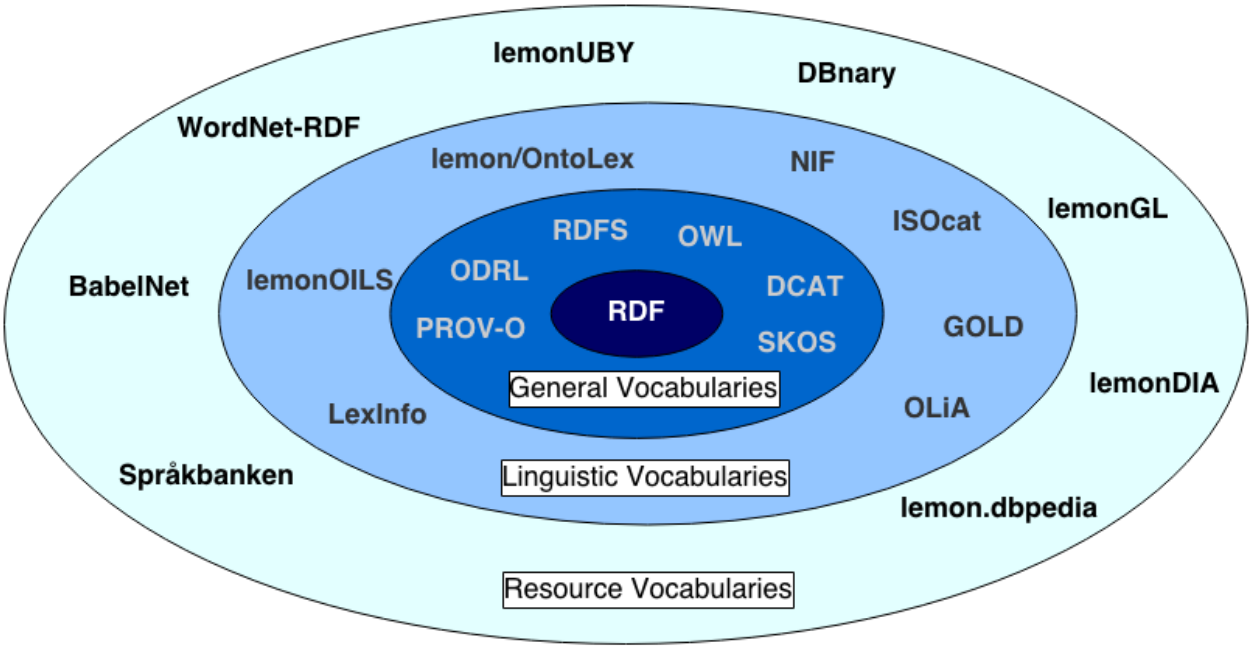
3 Identification of best practices and vocabularies for metadata for multilingual and multimedia web content

3.1 Vocabularies

As part of the work on the identification of best practices and vocabularies for description of metadata for multilingual and multimedia web content, we have identified relevant vocabularies to be used in the description of linguistic resources. We group the vocabularies along three layers built around the Resource Description Framework (RDF) as a core. We consider first so called **general vocabularies (3.1.1)**, which are

standardized by W3C and can be used not only for linguistic resources but also all kinds of resources. These vocabularies are recommended to be reused for any dataset.
Secondly, **linguistic vocabularies (3.1.2)** are vocabularies used in linguistics to model particular types of resources such as lexicons or corpora. These vocabularies are recommended to be used for datasets modelling linguistic information.
Finally, **resource-specific vocabularies (3.1.3)** are generally used only by a single resource and represent specific vocabulary relevant to their modelling. These vocabularies are specific for a particular type of resource and should be used when modelling that type of resource.
In this section, we only provide an overview of the relevant vocabularies. Section 2 shows how these vocabularies are used in the modelling of particular resources.

The following diagram depicts the different layers:



https://www.draw.io/#G0BwvuzIAhamr9TmlLb0ZMT2JoS2s

In the following, we provide a short description of these vocabularies:

3.1.1 General Vocabularies

- **RDFS[2]:** The RDF schema vocabulary covers description of general properties of data, such as domain and range of properties

[2] http://www.w3.org/2000/01/rdf-schema#

- **OWL**[3]**:** The Web Ontology Language covers axiomatization of data models and the description of ontologies
- **SKOS**[4]**:** The Simple Knowledge Organization System is used to model classifications, thesauri and other semantic networks
- **DCAT**[5]**:** The Data Catalogue Vocabulary can be used to describe the structure and metadata of datasets. The Vocabulary of Interlinked Datasets (**VoID)**[6] plays a similar role.
- **PROV-O**[7]**:** The Provenance Ontology describes the source and processes used to create data
- **ODRL**[8]**:** The Open Digital Rights Language describes how a dataset is licensed and how it can be accessed

3.1.2 Linguistic Vocabularies

- ***lemon***[9]**/OntoLex**: The Lexicon Model for Ontologies, developed by the OntoLex[10] community group, is an emerging standard for the representation of lexicon and other dictionary-like resources
- **NIF**[11]**:** The NLP Interchange Format provides the ability to add stand-off annotations to any text in RDF format
- **ISOcat**[12]**:** This is a collection of data categories used for linguistic description
- **OLiA**[13]: The Ontologies for Linguistic Annotation provide mappings between a number of linguistic annotation schemes
- **GOLD**[14]**:** The General Ontology for Linguistic Description contains a number of useful concepts for linguistic description
- **lemonOILS**[15]**:** The *lemon* Ontology for Interpreting Lexical Semantics allows a number of lexico-semantic phenomena to be added to OWL ontologies
- **LexInfo**[16]: LexInfo provides a large number of linguistic categories aligned with *lemon* and ISOcat

3.1.3 Resource-specific vocabularies

- **WordNet-RDF**[17] provides a number of terms specific to the modelling of WordNets

[3] http://www.w3.org/2002/07/owl#
[4] http://www.w3.org/2004/02/skos/core#
[5] http://www.w3.org/ns/dcat#
[6] http://www.w3.org/TR/void/
[7] http://www.w3.org/ns/prov#
[8] http://www.w3.org/ns/odrl/2/
[9] http://lemon-model.net/lemon#
[10] http://www.w3.org/community/ontolex/
[11] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
[12] http://www.isocat.org/datcat/
[13] http://purl.org/olia/olia.owl#
[14] http://purl.org/linguistics/gold/
[15] http://lemon-model.net/oils#
[16] http://www.lexinfo.net/ontology/2.0/lexinfo#

- **BabelNet**[18] is a large multilingual lexical resource derived from many sources, also available in *lemon* format
- **DBnary**[19] is a machine-readable version of the Wiktionary project
- A number of resources for Swedish are published at **Språkbanken**
- **lemonGL**[20] is an extension to *lemon* to support ideas from the Generative Lexicon
- **lemonDIA**[21] is used to describe Diachronic usage of *lemon* for Latin
- **lemon.dbpedia**[22] are a set of resources describing DBpedia in *lemon*
- **lemonUBY**[23] is the linked data version of the large-scale lexico-semantic resource Uby
- **lemon Translation**[24] is a module for representing explicit translations between senses of lexical entries expressed in different natural languages

3.2 Best Practices for Multilingual Linked Data

When the multilingual dimension of Linked Data is taken into account during the generation and publication process, some issues arise. The following set of patterns (or practises) identify possible ways to approach such issues. These are currently under analysis and discussion by the W3C BPMLOD community group[25] so the list is not final.

3.2.1 Practices for Naming

A key challenge for multilingual linked data is the naming of elements by means of URIs or IRIs. The following alternative naming schemes are possible:

- **Descriptive URIs**: Such as `http://example.org/Armenia`, these URIs are readable in a single language and are better supported by tools which display URIs directly to users. Furthermore, for many ontologies they are the only way of encoding language data without using labels. However, this pattern is not applicable to non-Latin languages, due to the fact that non-ASCII characters are represented percentage encoded, which hinders readability. Furthermore, for many domains it is not possible to encode sufficiently descriptive names within the space of a URI.
- **Opaque URIs:** Such as `http://example.org/5694bff3-7aae-4157-a237-0418dae17dc7`, these URIs represent internal identifiers or *globally unique identifiers* (GUID). Such URIs provide independence between the content and the language and can better handle changes in the label of a concept. Such URIs should also be preferred if they follow an existing ID scheme in an extant

[17] http://wordnet-rdf.princeton.edu
[18] http://babelnet.org/2.0/page/
[19] http://kaiko.getalp.org/about-dbnary/
[20] Not yet released, under development
[21] Not yet released, under development
[22] http://github.com/cunger/lemon.dbpedia/target/
[23] http://purl.org/olia/ubyCat.owl#
[24] http://purl.org/net/translation
[25] http://bpmlod.github.io/report/patterns/index.html

language resource. However, such opaque URIs are not human-readable and may be difficult to handle by developers.

- **Full IRIs:** Such as `http://օրինակ.օրգ/Հայաստան`, these are IRIs contain non-ASCII text. Many of the same issues apply as for descriptive URIs, but in addition there are issues related to the lack of tool support, security issues, particularly due to characters that resemble standard ASCII characters (spoofing), and issues with languages written from right-to-left. Finally, while most internet users are familiar with the Latin alphabet, special text input methods are required for many non-Latin alphabets that make such URIs difficult to handle for users unfamiliar with the language.
- **Path-only IRIs:** Such as `http://example.org/Հայաստան`, these IRIs include non-ASCII characters only in the path. This reduces security and tool risk as only the server hosting the data needs to understand non-ASCII characters, and makes it easier to fall back to percentage encoding if necessary.
- **Per-language Descriptive Identifiers:** Such as `http://en.example.org/Armenia`, `http://tr.example.org/Ermenistan`, a URI or IRI is set up for each language. This may also be done for practical reasons such as to split datasets from different language sources (e.g., DBpedia). Finally, it is important to distinguish between the use of a language code in the domain, which requires DNS configuration, and the language code in the path or file name, which enables content negotiation (see below)

In conclusion, **the use of descriptive URIs or opaque URIs is recommended**. However if IRIs are used it is preferable that an ASCII domain is still used (path-only IRIs). In any case, it cannot be assumed that descriptive URIs/IRIs are a means for encoding language data. Other techniques (e.g., labelling) have to be used to that end (see “practices for textual information” below).

3.2.2 Practices for Dereferencing

Multilingual data is often very large and as such the use of content negotiation to return only labels in the requested language by means of the HTTP `Accept-Language` parameter is recommended. Here, several options exist:

- **No Language Content Negotiation:** The multilingual labels are part of the data and thus all the data including the labels in all languages should be returned. All services should implement this as a fallback if no language is specified in content negotiation.
- **Language Content Negotiation:** Here the same content is provided but with only labels in the requested language(s). This can save bandwidth for some applications, but it is unclear if this provides real advantages over SPARQL queries or API-based access.
- **Language Content Redirection:** In this case the client is redirected to a URI containing the monolingual data. As such each URI still represents the same set of triples.

In general we regard language content negotiation as optional.

3.2.3 Practices for Textual Information

The representation of textual information in linked data is a key element. As such we have identified the following practices for including multilingual information in linked data:

- **Label Everything**: This pattern states that every single element in a resource should be labelled in as many languages as possible by means of the `rdfs:label` property. In general, this is preferable but some tools may support other labelling properties.
- **Multilingual values:** Data properties used in datasets may have natural language strings as their values. In this case it is necessary to include a language tag and if possible provide multilingual variants.
- **Untagged labels:** In addition to a language-tagged literal values, it is also possible to include an untagged literal value. This enables querying from applications where the language is not specified and increases usability of the SPARQL by allowing queries that do not take into account language tags to work. This can be considered an anti-pattern and is not generally recommended.
- **Break up longer descriptions:** A long text literal can be split up into a smaller annotations. For example the annotation that a person's job description is "Professor at Bielefeld University" could be split into two statements that their position is "Professor" and their place of work in "Bielefeld University". Such statements could also then be linked by means of the NIF vocabulary. The applicability of this pattern is clearly limited to cases where queries over separated part of otherwise long labels are feasible.
- **Provide full lexical description:** Each entity should be named by means of reference to some lexical entry expressed in *lemon*. By this, a higher level of linguistic description is achieved that can be exploited by applications requiring deeper linguistic information. This might represent unnecessary overhead for many applications.
- **Structured Literals:** In some cases. instead of providing a tagged literal it may be preferable to include an XMLLiteral. In this pattern it is still important to provide the language tag within the XML element. For example, it may be of interest to use the Internationalization Tag Set to create a literal such as:
  `<span xml:lang="de"> Universität <span translate="no"> Bielefeld </span> </span>`

In general the use of language tags is strongly recommended for even monolingual data. In addition, the use of vocabularies such as *lemon* or NIF may further support the reuse of multilingual data in a wide number of applications

In general, the use of `rdfs:label` is recommended as a baseline in all cases. If needed, richer descriptions can be provided exploiting models such as lemon, ontolex, SKOS, etc.

3.2.4 Practices for linking

When there are similar resources in multiple languages, it should be possible to provide links between them to enable multilingual use. We consider the following options:

- **Inter-lingual identity links:** Identity stating properties such as `owl:sameAs` can be used to establish links between equivalent resources in different languages. Such properties are supported by OWL reasoners but may lead to undesired results if the resources contain contradictory information, e.g. originating from a differing conceptualisation of a resource resulting from legitimate linguistic and cultural differences between speakers of different languages.
- **Inter-lingual soft links:** Alternatively, weaker properties such as `rdfs:seeAlso` can be used, but as such properties have no clear semantics, their value for applications is limited.
- **Linking to a common index:** Finally, instead of establishing links directly, it may be preferable to establish indirect links by mapping to a common ontology such as DBpedia or BabelNet. For example a Swedish resource may distinguish 'farfar' (paternal grandfather) and 'morfar' (maternal grandfather), both of which could be linked by hypernym relations to the English concept of 'grandfather' in WordNet or another semantic resource.

In general, **linking should be done at the level with the strongest semantics** that is justified. Clearly, this depends on how close the two multilingual resources are.

3.2.5 Identification of languages

Finally a key issue is the identification of languages. There are a number of standards of which we will outline here:

- **ISO 639-1[26]:** The two letter code defined by ISO 639-1 identify most of the 'commonly-spoken' languages. However, the coding fails to cover many languages with over a million speakers especially in Asia, while supporting several minor languages in Europe.
- **ISO 639-3:** The three-letter codes defined by ISO 639-3 allow for codes to be developed for all identified languages in the world. However, this is not backwards compatible with ISO 639-1 and does not cover many dialects.
- **ISO 639-1 with fallback to ISO 639-3:** In this case the two letter code should be used to identify the language if it exists, otherwise the three letter code should be used, e.g. 'en' for Standard English, 'sco' for Scottish English.
- **BCP-47[27]:** This Best Common Practice builds on the above mentioned standards but also incorporates ISO-15924 for identifying scripts and ISO-3166 for identifying geographical locations to model dialect. Furthermore, the use of IANA subtags is supported for representing information such as Romanization scheme and other variants. This standard adopts tags and subtags for representing languages and variants; for instance, German is encoded as "de", French as "fr"

[26] ISO 639 is the International Standard for language codes. It is composed of six different parts. For more details, see http://www.iso.org/iso/home/standards/language_codes.htm
[27] http://www.rfc-editor.org/bcp/bcp47.txt

but the “Chinese, Mandarin, Simplified script, as used in China” is encoded with the label “zh-cmn-Hans-CN”.

- **LoC URIs:** The Library of Congress maintains RDF descriptions for language codes in ISO 639-1 but these are limited.[28]
- **LexVo URIs:** The LexVo project maintains RDF descriptions for all ISO 639-1 and 3 codes[29] but does not support the specification of script or regional dialect.

In general, **BCP-47 is the recommended way to specify the language**, and while it may be advantageous in the future to switch to URIs which can be dereferenced for full RDF descriptions, there is currently no complete and well supported service that enables this.

3.3 DataID

The constantly growing amount of Linked Open Data (LOD) datasets has triggered the need for rich metadata descriptions, enabling users to discover, understand and process the available data. This metadata is often created, maintained and stored in diverse data repositories featuring disparate data models that are often unable to provide the metadata necessary to automatically process the datasets described.

The importance of describing datasets using VoID is well established, but there is still a lack of important metadata which is not described, for example license and provenance. The DataID data model is proposed as a best-practice for LOD datasets description in that it provides a uniform way to describe general metadata of datasets in RDF format. Thus, describing a dataset using the DataID model allows to determine what category it belongs to, what other datasets it is linked to, where example resources can be found, who published the dataset under which license and much more. Likewise DataID tackles three important aspects:

- **PROVENANCE:** A crucial aspect of data which is needed to assess correctness and completeness of the data conversion, as well as the trustworthiness of the data source.
- **LICENSING:** Machine-readable licensing information is crucial as it provides the possibility to automatically process and publish only data that explicitly allow these actions.
- **ACCESS:** Finally, publishing and maintaining this kind of metadata together with the data itself serves as LOD-compatible documentation benefitting the potential user of the data as well as the creator by making it discoverable and crawlable. Thus, DataID uses vocabularies for dataset description based on DCAT, VoID, DCTerms, Prov-O and several extensions.

The three aspects cited above can be seen with details in the following description of Figure 1. As can be seen, the DataID model exploit 7 main classes from existing

[28] http://id.loc.gov/vocabulary/iso639-1.html
[29] http://www.lexvo.org/page/iso639-3/eng

vocabularies. An essential description of each class of the model was made in order to clarify the overall structure of the DataID model:

- **Dataset class:** `void:Dataset` defines the Dataset class whose properties are particular to RDF datasets like `void:triples` and `void:sparqlEndpoint`, as well as criteria on how to use other vocabularies such **DCTERMS**. As mentioned above, the DataID data model uses VoID extensively and includes the RDF specific properties. It also includes the property `void:subset` to introduce descriptions of parts of datasets.This is especially important to describe monolingual subsets of multilingual datasets. **DCAT** also defines a dataset class, `dcat:Dataset`. Like **VoID**, it also uses **DCTERMS** properties for general metadata, thus enabling us to merge the dataset concept of both vocabularies into `datid:Dataset`.
- **Linkset class:** `void:Linkset` was adopted as well, enabling the description of content and number of links between different datasets. By using linksets, visualizations that show connections between datasets, like the ubiquitous LOD cloud[30] diagram, could be easily realized without having to directly access and process the data.
- **License class:** `datid:licenseName` contains the name of the license linked by the `dcterms:license` property, with the result that the name of the license can be easily retrieved without querying another resource on a different server.

[30] http://lod-cloud.net/

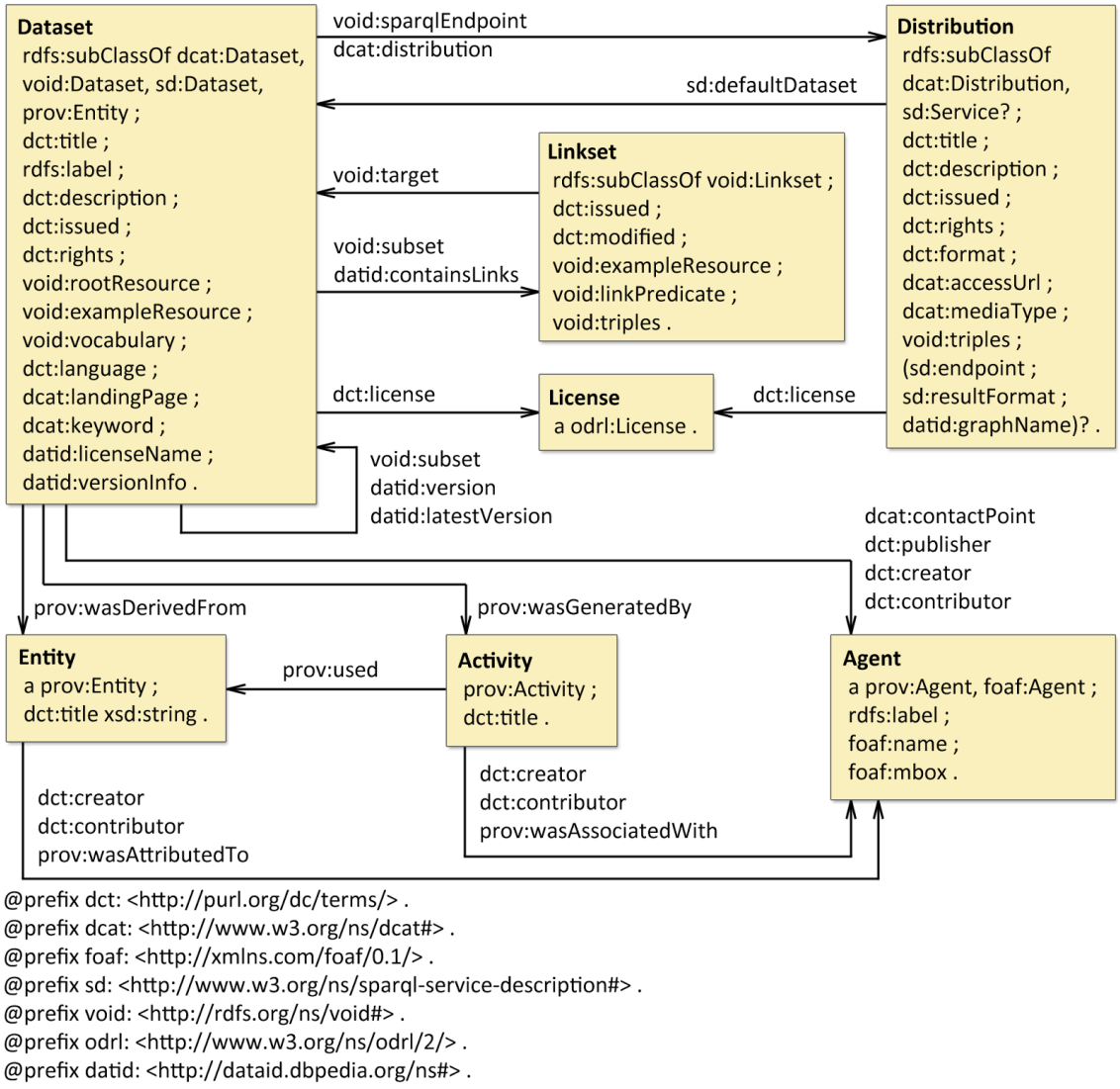


Figure 1: Overview of DataID data model

- **Distribution class:** `dcat:Distribution is used` for further description of directly accessible serializations of the data itself. This concept is crucial to be able to automatically retrieve and use the data described in the DataID, simplifying, for example, data analysis. To account for SPARQL-endpoints as special type of distributions, the SPARQL Service Description (SD)[31] vocabulary was used to differentiate it from file-based distributions. Properties used include sd:endpoint to link to the endpoint's URL, `sd:resultFormat` to annotate the format of the results, as well as the class `sd:Service` as a type of distribution. Because describing the name of the graph that contains the dataset's data in SD would imply creating a `sd:Graph` resource and thus create too much overhead, `datid:graphName` was defined. It contains the name of the relevant graph as a literal string.

[31] http://www.w3.org/TR/sparql11-service-description/

- **Entity, Agent and Activity Classes** from the **Provenance Ontology** (**Prov-O**) are included together with their respective properties in order to capture a complete, fine-grained provenance chain.

Concerning the deployment of the DataID file, it has to be determined and implemented by its users. To ease adoption and in order to be as inclusive as possible, we are following the robots.txt convention. This widely implemented de-facto standard consists of a file robots.txt in the top-level directory of a web server that contains a number of constraints regarding which URLs of the web site are forbidden from being visited by programs automatically crawling it. Similarly, we propose DataID files to be put in the top-level directory of the web server. The format must be Turtle, which is the RDF serialization featuring the best compromise between readability and file size. The file should simply be called dataid.ttl. For a DBpedia scenario, next to the http://dbpedia.org/robots.txt that explicitly excludes certain directories from being crawled, the http://dbpedia.org/dataid.ttl explicitly states where DBpedia datasets can be found as well as their content and relevant metadata. This best-practice allows publishers to aggregate descriptions of all hosted datasets in one place and also enables users to easily discover and access these datasets.

The generation of DataID files might be an overwhelming task. For small datasets, a tool was created in order to help users to create DataID files. The DataID generator[32] can handle the generation of DataID files of small or mid-size datasets. In cases of large datasets we recommend the generation using scripts and afterwards use the DataID validator[33] .

3.4 OWL Metamodel for Language Resources

With the aim of converting data/metadata of Language Resources into the cloud of Linguistic Linked Data, an OWL metamodel for Language Resources is currently being studied and developed in the LD4LT group[34]. This is based on the inputs and previous experiences of well-established LRs communities such as Meta-Share, CLARIN, LREMap, etc. More details on each of these repositories and a comparative study can be found in Section 3. The approach followed in developing the OWL metamodel is bottom up, meaning that initially the different representation schemes will be analysed and converted into an OWL model and, in a future step, a minimal common model will be extracted to allow interoperability among LR repositories.
For the time being, the work at LD4LT has been initiated for the Meta-Share metamodel. In particular, the core of the Meta-Share model has been already analysed and mapped to common vocabularies such as DCAT wherever possible. An extensive analysis of the

[32] http://dataid.dbpedia.org/
[33] http://dataid.dbpedia.org/validator/
[34] https://www.w3.org/community/ld4lt/wiki/Main_Page#OWL_Metamodel_for_Language_Resources

semantic representation of their licenses has been also carried out (see below). The analysis of the LREMap metamodel is expected to start shortly.

3.5 License Ontology

The purpose of this section is to present the licensing information about a resource and published data in general. A license is a document which regulates the permission to access, modify, copy and redistribute material.

Depending on the specific needs, publishers might be satisfied with referencing existing general-purpose licenses (goal of section 1.5.1) or they might be forced to describe the license making use of more complex ontologies such as ODRL (section 1.5.2).

Finally, in section 1.5.3 we describe the licenses introduced in the Metashare model in order to illustrate the different levels of detail and fine-grained distinctions a publisher might want to consider when publishing a resource in RDF.

3.5.1 Simplest recommended practice for licensing language resources

Declaring the license for every published resource is a needed practice if resources are to be reused. Declaring a license in RDF is simple. The standard property to declare the license of a resource is the Dublin Core `dct:license`: `http://purl.org/dc/terms/license`

In some cases, the parent element is also considered `(dct:rights)`; or sometimes even some other properties like `cc:rights` (in the namespace of the Creative Commons REL[35]).

The license element is used to reference either an external document or a well-known license. An example of the former could be Microsoft license agreement for their resources in the linguistic portal[36], an example of the latter might be a Creative Commons attribution license[37] whose URI is widely known.

The understanding of the licensing terms by following this recommendation grants that humans become aware of the licenses, and that they have an easy access to the legal text.

Linguistic Linked Data resources, which shall always be identified by a URI, should be the subject of the RDF triple declaring the license. A good practice is declaring the resource to be a unit of distribution (for example dcat:Distribution) and generating as many distributions as differently licensed parts of the resource exist.

3.5.2 Complex recommended practice for licensing language resources

If the license declaration described in the previous Section does not suffice, a richer description should be provided. This is the case if the entity publishing the resource

[35] http://creativecommons.org/ns
[36] http://www.microsoft.com/Language/es-es/LicenseAgreement.aspx
[37] http://creativecommons.org/licenses/by/4.0/

(being the rightsholder or acting on behalf of the rightsholder) has specific interest in expressing this license as RDF. Licenses expressed in RDF allow deploying Digital Rights Management systems, eases the rights clearance process and allows a better data structure for archiving, so that machines can also unambiguously understand the rights.

If one of the pre-defined licenses suffices, the RDF version might be found as Linked Data within the following dataset: `http://datahub.io/es/dataset/rdflicense`
However, additional work has to be done if specific licenses (not those reduced to the set of well-known licenses) are to be used. In this case, Rights Expression Languages and Policy Language permit expressing the details of the licenses. This is the case of ODRL 2.0[38], which provides a generic language for expressing permissions, prohibitions and obligations.
An extension of ODRL2.0 (Linked Data Rights) has been proposed to express licensing information for linked datasets, e.g.

`http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/`

This vocabulary provides the elements for producing rights expressions for Linked Data. As an example of use of ODRL for representing actual licenses, one of the Metashare licenses is shown here (see in Figure 2 the "META-SHARE NonCommercial NoRedistribution NoDerivatives For-a-Fee Licence", whose text is online at `http://www.meta-net.eu/meta-share/licenses`)

```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix odrl:  <http://www.w3.org/ns/odrl/2/> .

<http://example.com/nc-nored-nd-ff> a odrl:Policy ;
    rdfs:label      "NC-NoReD-ND-FF" ;
    rdfs:comment    "MetaShare NonCommercial, No Redistribution, No Derivatives, for a fee.
                     Perpetual, worldwide, allowing no redistribution of the original. "@en ;
    rdfs:seeAlso    <http://www.meta-net.eu/meta-share....pdf> ;
    odrl:permission  [ a            odrl:Permission ;
                       odrl:action  odrl:reproduce;
                       odrl:duty    [ a            odrl:Duty ;
                                      odrl:action  odrl:pay ;
                                      odrl:target  "XXX EUR"
                                    ]
                     ] ;
    odrl:prohibition  [ a           odrl:Prohibition ;
                        odrl:action odrl:commercialize, odrl:distribute, odrl:derive
                      ] .
```

Figure 2. Example of Metashare license expressed in RDF with ODRL

This compact expression shows how the key elements are captured: the prohibition of the commercialization, distribution or derivation; its access only granted upon payment, etc.

[38] http://www.w3.org/community/odrl/two/model/

4 Development of guidelines and models for Linguistic Linked Data generation, publication and exploitation

The goal of this task is the development of guidelines that support the entire lifecycle of linguistic linked data resources, starting from i) the **modelling and generation of linked data resources,** over ii) the **publication** of these resources over the web, through to the iii) **exploitation** of these resources in content analytic tasks. In this first phase, we have focused on the development for **guidelines for the generation and publication of a number of frequent and relevant types of linguistic resources,** including wordnet, bilingual dictionaries, terminological resources (in TBX format), BabelNet as well as natural language processing services. All these guidelines are living documents that are developed by the relevant communities driven by the LIDER project. As these guidelines are living specifications, in this deliverable we concentrate on stating the motivation of the LIDER project for the development of this particular set of guidelines and we describe the community process and involved communities. A snapshot of the current state of the guidelines can be found in the Appendix to this document.

4.1 Guidelines for Converting WordNets to Linked Data

WordNet is a large lexical database of English nouns, verbs, adjectives and adverbs. Word forms are grouped into more than 117,000 sets of (roughly) synonymous word forms, so called *synsets*. These are interconnected by bidirectional arcs that stand for lexical (sense-sense) and semantic (synset-synset) relations, including hyper/hyponymy (*tree-oak*), meronymy (*tree-branch*), antonymy (*long-short*) and various entailment relations (*buy-pay, show-see, untie- tie*).

Rationale: The main reason for looking into WordNets as a resource to be converted into Linguistic Linked Data is that WordNets are widely and frequently used within computational linguistics for many tasks that can benefit from lexical background knowledge. We focused in particular on Princeton WordNet as it definitely has the potential to become a “de facto” hub for the emerging Linguistic Linked Data cloud (LLOD). A more detailed motivation for focusing our efforts on converting Princeton WordNet can be found in the following paper published at LREC.

Community Involvement: The wider community was involved as described below. John McCrae spent two months at Princeton working on the publication of Princeton WordNet together with the group of Dr. Christiane Fellbaum. The publication was later discussed at relevant conferences in the field, i.e. the Global WordNet Conference 2014 in Estonia, as well as the International Conference on Lexical Resources and Evaluation (LREC) in Reykjavík. The resulting dataset was further disseminated among relevant mailing lists and feedback was gathered from the wider community and implemented wherever it was found reasonable.

The guidelines that were developed from our experience with converting the Princeton WordNet into Linguistic Linked Data can be found here:

https://www.w3.org/community/bpmlod/wiki/Converting_WordNets_to_Linked_Data

The dataset resulting from the conversion of Princeton WordNet can be found here:

http://wordnet-rdf.princeton.edu/

4.2 Guidelines for Linguistic Linked Data Generation: Multilingual Knowledge Bases

Rationale: BabelNet represents one of the biggest multilingual lexico-semantic knowledge bases available to the community. The conversion of BabelNet into a lemon-based model made it possible to have a SPARQL endpoint which seamlessly enables users to query BabelNet using the LD paradigm.
Beneficiaries include thus not only academics but also companies and business stakeholders who can now exploit this resource as one of their building blocks for commercial products. This guideline was developed as part of the LIDER project so as to act as a reference model for all those who wish to convert another multilingual resource into linked data.

Community Involvement: During the process, people coming from different areas and universities were involved, each with a specific role in the project. The conversion involved different skills as well, such as system administrative competence, linguistic expertise but also specialists in the resource peculiarities. The conversion also resulted in a paper publication at LREC "Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0" and ongoing collaboration is happening between several university research groups, among which the Ontology Engineering Group at UPM (Madrid) and the Unit for Natural Language Processing at INSIGHT (National University of Ireland, Galway).
The conversion also caught the attention of many companies which are demonstrating increasing interest, effectively bringing the academic and the commercial worlds closer than before in this area.

The current state of the guidelines for the generation of multilingual knowledge bases, in particular for the case of BabelNet, can be found here:

http://bpmlod.github.io/report/multilingual-dictionaries/BabelNet/

A SPARQL endpoint and its LD interface are available at http://babelnet.org:8084/sparql/ and http://babelnet.org/2.0/page/ respectively.

4.3 Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries

Bilingual dictionaries are a type of dictionaries used to translate words or phrases from one language to another. They can be unidirectional or bidirectional, allowing translation, in the latter case, to and from both languages. In addition to the translation, a bilingual dictionary usually indicates the part of speech, gender, verb type, declination model and other grammatical properties to help a non-native speaker use the word. We are interested in bilingual dictionaries that have their data in a machine-processable format, no matter whether it is stored locally or is accessible on the Web (e.g., for download). We assume that the data is represented in a structured or semi-structured way (e.g., relational database, xml, csv, etc.).

Rationale: There is a number of bilingual and multilingual electronic dictionaries developed by different communities in isolation, following non-standard formats and hidden behind proprietary APIs in many cases. This hampers re-usability and interoperability at a Web scale. This is precisely what a LD version of such resources is able to attain. Such bilingual/multilingual dictionaries as LD will be potential sources of background knowledge for Machine Translation, cross-lingual ontology mapping, cross-lingual Question Answering, etc.

Community Involvement: At the model level, the Ontolex community group is carrying out a study for representing translations and other linguistic variations as linked data. This guideline constitutes a practical realisation of some of the modelling ideas discussed in Ontolex. The guidelines proper have been developed and published in the context of the BPMLOD community group and are currently receiving feedback from their members. Important stakeholders such as the Wikimedia Foundation and Kernerman Dictionaries have shown interest in the guidelines as well as in our first demonstrator based on Apertium (an open source machine translation system). Such an RDF version of the Apertium dictionaries[39] has been developed by Jorge Gracia and Asunción Gómez-Perez (UPM) jointly with Marta Villegas and Núria Bel (Universitat Pompeu Fabra). This work has been disseminated through the Apertium mailing lists where it has received very valuable feedback.

The most recent version of the guidelines can be found here:

http://bpmlod.github.io/report/bilingual-dictionaries/index.html

A snapshot of the guidelines can be found in the Appendix.

[39] http://linguistic.linkeddata.es/apertium/

4.4 Guidelines for Converting TBX into Linked Data

TBX is an open standard for sharing of terminological data that has been published by the Localization Industry Standards Association (LISA) (see here). The standard is identical to ISO standard 30042.

Rationale: During our roadmapping activities it became clear that terminological information is crucial for many applications, either to ensure terminological consistency or as a way to include background knowledge about a domain into content analytic tasks. The prevalent standard for sharing terminologies is the Term Base eXchange (TBX) language, an XML format standardized by ISO under standard 30042. It is the most commonly used exchange format. So far, an RDF / linked data version for TBX that exploits standard vocabularies was not available. Having an RDF version of terminological data supports easier integration of data as well as more flexible and easier querying using SPARQL.

Community Involvement: The TBX2RDF Converter was developed in interaction with a number of communities. First of all, industrial stakeholders have been involved in the process, providing requirements and guidance on the development of the converter. In particular, the TILDE company (http://www.tilde.com/) has been directly involved in the process. We have also involved the community around ISO TC37 (Alan Melby, Sue Ellen Wright) via Arle Lommel (DFKI). Interaction with the TC37 community will need to be intensified in the next project phase. The document describing the guidelines for converting TBX to RDF have been produced as part of the activities of the Best Practice for Multilingual Linked Open Data (BPMLOD) community group and have been discussed intensely in the context of this group. We have further organized a hackhaton on the topic of converting TBX data to RDF at the MLODE workshop in Leipzig, gathering an additional community of developers on the project and gathering additional feedback.

The most recent version of the guidelines can be found here:

http://www.w3.org/community/bpmlod/wiki/Converting_TBX_to_RDF

A snapshot of the guidelines can be found in the Appendix.

4.5 Guidelines for NIF-based NLP Services

Rationale: Many NLP services and tools exist nowadays, but they are largely fragmented according to framework (GATE vs. OpenNLP vs. UIMA …) and as such do not interoperate well. The goal of the LIDER project is to contribute to the emergence of a unified ecosystem of NLP services that seamlessly interoperate and can be composed into more complex workflows. The guidelines provide recommendations on how to set up NLP services as NIF-compliant RESTful web services, taking a NIF document as input and producing a NIF document as output. As proof-of-concept, we demonstrate the application of the guidelines to implement NIF-based wrappers for the Stanford POS

tagger and the Stanford Parser, showing how the tools can be concatenated in tool chain using curl.

Community Involvement: The guidelines have been developed in joint collaboration with the NLP2RDF Community Project.

The most recent version of the guidelines can be found here:

https://www.w3.org/community/bpmlod/wiki/NIF_Web_Services

A snapshot of the guidelines can be found in the Appendix.

5 Development of guidelines for LLD-aware NLP services

Task 2.3 focuses on the use of Linguistic Linked Data in content analytics, in particular by means of "LLD-aware NLP services". Such services support content analytics by exploiting Linguistic Linked Data resources on the Web. In particular, a key goal here is the **discovery**, **delivery** and **extraction** of language resources from the Web. The guidelines will describe how such systems can seamlessly download these resources, either as a full resource or only required slices of the resource. Finally, the guidelines will describe how these resources can be quickly converted into a form that can be used in an existing content analytics process.
The focus is thus clearly on providing guidelines to publishers of linguistic linked datasets to facilitate the discovery and exploitation by NLP services, rather than on providing guidelines on how to implement the LLOD-aware services proper.

As a first step in producing such guidelines, in the first phase of the LIDER project we have concentrated on analysing the current state of existing repositories as a basis to i) provide a clear set of recommendations to parties that host and maintain such repositories, and ii) as a basis to develop guidelines for how LLOD-aware services can interact with such repositories to discover and extract the data they need to perform a certain task.

5.1 Overview of repositories

The development of guidelines for LLD-aware and natural language processing services is founded on a comparative analysis of already existing data repositories for linguistic data. The overview has been carried out for the following five repositories of linguistic data:

- Datahub[40],
- Multilingual Europe Technology Alliance (META-SHARE[41]),

[40] http://datahub.io/
[41] http://metashare.dkfi.de/

- Common Language Resources and Technology Infrastructure (Clarin[42]),
- Language Resources Evaluation Map (LRE Map[43]), and
- Linguistic Data Consortium **(**LDC[44]).

The choice of these repositories is motivated by the considerable number of users they have, the wide range of linguistic data they cover, as well as their impact on the linguistic data community. The aim is not only to provide a review of the linguistic data portals but also to identify existing flaws that will be taken into consideration in ongoing LIDER project activities.

In the following two sections, an overview of the linguistic data repositories (3.1.1), a comparison of selected repository features (3.1.2) and a summary of recommendations for meeting the European Commission's Open Data Strategy (3.1.3) will be given.

5.1.1 Existing Data Repositories

Datahub is a platform developed by the Open Knowledge Foundation that enables users to upload, group and search open data. It is built on the Comprehensive Knowledge Archive Network (CKAN) software and provides metadata about (Linked Data) datasets. Datahub hosts data not restricted to any particular domain and enables both data providers and users to edit dataset entries. By design, the metadata provided for most of the datasets is flat and simple. On the one hand, this facilitates the ease of upload and addition of datasets for providers, but, on the other hand, it reduces the usefulness of the metadata and the accessibility of the data itself.
However, the most significant problem of the dataset management at Datahub is the absence of detailed provenance information. A simple activity stream captures who applied changes in the dataset entry. Datasets in RDF format for example do not reference the data source they were derived from in a proper way. As a consequence, Datahub is incapable of adequately describing datasets with multiple source datasets and files such as the well-known DBpedia.
Another linguistic data repository is **META-SHARE,** which is part of the Multilingual Europe Technology Alliance (META) and aims at providing quality language resources. Focusing on the processing of the metadata, a strictly provider-driven account is taken. META-SHARE assumes a high quality of the data it hosts, because only scientific institutions are allowed to add datasets. Once the data is integrated into the repository, no further data validation is conducted. For the data providers, however, it is often infeasible to update the metadata in regular intervals and there is no issue reporting by user requests. This contributes to a rather static approach to data storage and leads to an unbalanced data repository favouring data preservation that contributes little to effective data reuse.
A quick and structured access to information on language resources is provided by the **Language Resources Evaluation Map** (LRE Map). It originated in 2010 at the LREC

[42] http://clarin.eu/
[43] http://clarin.eu/
[44] https://ldc.upenn.edu/

conference where all contributing authors were asked to fill in a form asking for information about the language resources they used. In the following years, authors from other conferences joined this procedure as well, so that a matrix of nearly 4,000 language resources emerged. A faceted search functionality is realized through various metadata values that can be multiply selected. However, the LRE Map does not host any of the resources it lists in the catalogue and provenance only goes as far as referencing the project page of the dataset (not the download links). Therefore the LRE map basically represents a collection of metadata that is restricted to a selected group of data providers and offers only a display of language resource names and categories to the user.

A complex and sustainable repository network for digital language data is implemented by the **Common Language Resources and Technology Infrastructure** (CLARIN). This repository network is distributed among selected research centers in the humanities and social sciences across different countries. The realization of metadata compilation and storage is based on gathering metadata descriptions which are used to set up a so called *Component Metadata Instance (CDMI)* that creates a CDMI metadata file for each language resource. Metadata categories are fixed and bound to ISOcat categories but editable by every CLARIN member. By this, greater dynamics within the metadata maintenance is assured. Furthermore, datasets can be downloaded and run on private computers or explored online. However, the data is neither in Linked Data format nor openly accessible to everyone.

One of the most important repositories for language resources is maintained by the **Linguistic Data Consortium** (LDC), which is run by the University of Pennsylvania. Among textual resources like corpora and lexical language sources, other valuable language materials such as audio and video files are supplied as well. Anyone who compiled a linguistic dataset is able to publish the resources using a corpus submission form. Each publication proposal is checked for completeness and errors in collaboration with LDC staff members. That way, the LDC assures a high quality of all datasets provided. On the downside, it has to be mentioned that Linked Open Data formats are neither required nor supported by the LDC. The contrary is the case: all datasets are bound to closed licenses and subject to fees for both LDC members paying less and for non-members charged full prices. As a consequence, the high quality of the provided datasets is repressed by the commercial data supply. Within the context of the Semantic Web this only adds obstacles that need to be overcome in order to make qualitative data freely accessible to everyone.

5.1.2 Comparison of Data Repository Features

As a starting point for our analysis, we considered the Open Data Strategy published by the European Commission in 2011, which has identified the following shortcomings (recited here for the sake of convenience):

I. A lack of information that certain data actually exists and is available;
II. A lack of clarity of which public authority holds the data;
III. A lack of clarity about the terms of re-use;

IV. Data which is made available only in formats that are difficult or expensive to use;
V. Complicated licensing procedures or prohibitive fees;
VI. Exclusive re-use agreements with one commercial actor or re-use restricted to a government-owned company.

Given these barriers, it becomes obvious that no universal remedy exists. The presentation of the five resource repositories above displays the diversity of the implementations adopted. Taking the data producer, the data user and the dataset itself as essential aspects for data repository setup, a detailed comparison of repository features is applied to the mentioned repositories (cf. Table 1) along five dimensions: i) repository content discovery, ii) accessibility, iii) data contribution iv) provenance constitution, and v) metadata processing.

- **Repository content discovery:** The most basic feature concerns the assessment of the *repository content*. A user visiting a website for the first time primarily wants to know what data is offered and if it corresponds to the data s/he is looking for. All repositories with the exception of Datahub identify themselves as domain-specific regarding the linguistic domain. With the certainty that the desired data is available, the user will go on to search more specifically for datasets. Therefore, an understanding of the structure of the repository content is necessary. The more complex and confusing the repository design is, the sooner a user will get frustrated by the search procedure and is likely to leave the repository unsatisfied. Due to its small size, an overview of LDC is straightforward to produce. LRE Map offers free-text search and faceted-browsing, which decreases the effort to find datasets. Since Datahub contains large amounts of datasets, a quick exploration is not feasible. This limitation is somewhat ameliorated thanks to group and tag facilities that help structure the dataset metadata. META-SHARE, however, being part of the more complex META-NET project, requires more examination of the repository internal data organisation. The most opaque repository structure is exhibited by CLARIN, because the repository content is visible to registered members only and the membership involves a three step authorization process in addition to the complex repository structure. Once the structure is understood, all five repositories allow for a domain-specific search via fixed linguistic categories.
- **Accessibility:** As soon as the user has found a relevant dataset, the second feature of *accessibility* becomes crucial. Ideally every dataset provides Linked Open Data or is at least licensed to be free, open and reusable for everyone. But this only applies to Datahub. The other repositories supply open and closed datasets (META-SHARE, LRE Map) or closed datasets only (CLARIN, LDC). Even if datasets are open in terms of licenses, all repositories with the exception of Datahub diminish the accessibility of these through various membership restrictions.

- **Data Contribution:** Assuming that the user has successfully been able to download the dataset of interest and used it already for research, questions of *data contribution* arise next. If mistakes were found in the metadata, only Datahub would allow for a direct editing of the metadata entry. The four remaining repositories reserve editing rights for the data providers or repository members only. Equally, an unrestricted integration of datasets into the repository has been realized in Datahub but not in META-SHARE, LRE Map and CLARIN. The latter ones restrict the upload and editing of datasets to registered members only. LDC gives everyone the possibility to submit own datasets. However, it strictly controls and adjusts it to LDC standards. Leaving the editing of metadata and dataset information to data providers or authorized persons only leads to static dataset descriptions that might contain unnoticed mistakes and withholds the community from taking over maintenance tasks.
- **Provenance Constitution:** A central problem within the domain of linguistic data deals with the issue of *provenance constitution*. With linguistics being an empirical research area, data usage demands for an explicit provenance chain. This includes information stating the source as well as the origin of derived datasets. A great variety of the provenance information provided is observable: Datahub and LRE Map merely state the resource URL or the web page that is supposed to host the dataset, forcing the user to collect all necessary provenance information from there. META-SHARE and LDC leave the user with an indication of the data sources and creators. A statement on how CLARIN is treating data provenance is not possible here, because we have no access to the repositories.
- **Metadata Processing:** The last repository feature is related to *metadata processing* and affects data users and providers alike. The repositories differ in the number of metadata elements, with the specific number of elements varying between 1 to 16 core elements. Thereby, META-SHARE compiles more metadata than is actually displayed with the dataset entry, including as only metadata the creation date of the dataset. With the exception of Datahub, metadata maintenance is done manually by the data providers or authorized members in every repository. Manual curation of technical information often results in inaccurate and outdated metadata due to the lack of resources. A way to tackle this problem could be the implementation of an automatic generation of metadata via dataset inspection and link analysis in addition to the manually edited metadata. That way, the accuracy of the technical metadata is not only assured but also facilitates the effort of dataset upload for the data providers.

After having gained insight into the various data repository constructions, three different kinds of data repositories can be identified:

(1) User-Centred (Datahub),
(2) Provider-Centred (META-SHARE, LRE Map, CLARIN),
(3) Data-centred (LDC).

Concluding, the presented repositories reveal an unbalanced emphasis on either the users, the providers, or the content of the linguistic resources. This results in the deficiencies outlined so far. Taking this repository survey as a basis, it is proposed to develop guidelines for LLD-aware NLP services that consider the EU Open Data Strategy and result in linguistic linked open data repositories which are easy and freely accessible to users, supporting data submission for data providers through an automatic metadata retrieval and assuring high quality datasets via dynamic repository structures cared for by the whole linguistic community.

5.1.3 Recommendations for Data Repositories

To summarize, the comparison of the selected linguistic data repositories reveals that none of them breaks down the six barriers defined in the Open Data Strategy of the European Commission. An attempt to tackle these is made by the Open Knowledge Foundation's Working Group on Open Data in Linguistics[45]. The community effort resulted so far in the development of a linguistic linked open data (LLOD) cloud that is built upon the linguistic resources hosted in the Open Linguistics Working Group (OWLG) on datahub.io[46]. What is more, these datasets are also embedded in the LOD cloud under the new category "Linguistics", given that they are tagged with "llod" and "lod" in the OWLG group in Datahub. The LLOD cloud as well as the import of its datasets into the LOD cloud can be seen as crucial steps toward the provision of "information that certain data actually exists and is available" (cf. first barrier above). With regard to these recent developments and the other barriers for open data, the following recommendations are given to the repository hosts which are highly advised to make associated adjustments.

5.1.3.1Datahub

Datahub is the only repository that consistently provides open and linked data. It could contribute to this even more by structuring the vast amount of datasets by offering strict domain labels rather than allowing arbitrary tagging of datasets. At the moment it is impossible to find all linguistic resources in Datahub because datasets are tagged with various labels such as "linguistic", "linguistics", "language", "corpus", "lexicon", "word list" and many more depending on the label the data provider comes up with. Given that the OWLG community is already generating the LLOD cloud overview on the basis of Datahub dataset entries, this could be supported by drawing the linguistic data providers' attention to this community effort and encourage them to also put it into the OWLG organization group and tag it with "llod" and "lod". Further improvement potential can be seen in the metadata curation. A set of obligatory metadata information fields are recommended, which should cover provenance and licensing information.

[45] http://linguistics.okfn.org/

[46] All resources can be seen here http://datahub.io/organization/owlg.

5.1.3.2META-SHARE

Within the linguistic and language technology communities META-SHARE is one of the well-established repositories for language resources. Its impact is however reduced within the semantic web research field due to the lack of resources provided as linked data. It is therefore strongly recommended that META-SHARE encourages the large amount of data providers they are connected with to transfer their datasets also into RDF. This should be ideally accompanied by a promotion to publish resources under an open license. Data providers who are in possession of a language resource in RDF format and are willing to provide these as publicly open and reusable data could be also referred to the OWLG group in Datahub and be advised to additionally publish their linked open dataset there with the “llod” and “lod” tags. Thereby, the central linked open data clouds would be enriched with META-SHARE’s high-quality datasets.

5.1.3.3LRE Map

The “Open Resource Infrastructure”, which the LRE community has developed, holds a significant number of nearly 4.000 resource entries. But taking a closer look at those reveals that only 869 of them are declared as “freely available”. Still, reaching the data is not easy, because no dataset owner details are given and less than half of these resources are provided with a resource URL from which the data can be directly downloaded. Furthermore, linked data resources are not explicitly declared. Only 134 datasets, which can be assumed to be in a linked data format, are under the resource type “ontology” and it is not clear if more linked datasets can be found within the other types.

Given that the LRE Map is in the valuable position to gather data at various events directly from the data providers, the community is advised to take this advantage and move beyond a mere representation of the dataset survey they hand to the data providers and consider the following adjustments:

- Make resources accessible by asking the resource owners to provide the URL for the dataset, or if not available a contact email address
- Open up the resources for real re-use by offering open licenses under the “availability” metadata category
- Promoting the idea of publishing the resources at datahub.io with the “llod” and “lod” tags
- Declare datasets in linked data format more explicitly to make them easier to find.

5.1.3.4CLARIN

Within the CLARIN Virtual Language Observatory nearly 700.000 language resources can be found. These can be looked up by using a number of search categories such as “language”, “resource type”, “format” and “data provider”. Extracting those datasets which are in a linked data format is, however, not possible since no linked data format is provided for search under the “format” category. It is therefore suggested to add at least the common linked data format

RDF in order to enable a quick access to all linguistic linked data datasets. A clear overview of which resources are openly available is not provided as well. The openness of each resource stays with the resource owners with the result that the conditions of data re-use have to be looked up for each dataset separately. Even though the data files of all open access datasets are directly given and downloadable, it is recommended to offer a search modality that allows for filtering out all freely reusable resources at once.

5.1.3.5LDC

The Language Data Consortium being founded in 1992 forms the oldest linguistic data repository and offers a wide variety of high quality language resources. Apart from not declaring or promoting any resources in linked data formats, the main obstacle of this repository is that it fully applies to the strongest barrier of the EU's Open Data Strategy: "complicated licensing procedures or prohibitive fees". Free and open access to the data is impossible to non-members. It is therefore strongly recommended to the LDC to at least consider providing some of the data it holds under an open license, which would also be the first step to enable other researchers to convert these datasets into a linked data format.

5.2 Harmonization of Repository Metadata

The different metadata repositories analyzed have their proprietary data models, which renders interoperability and uniform access to and discovery of resources on the basis of the metadata available from different repositories extremely difficult if not impossible. As a proof-of-concept of the discovery and querying functionality that would become possible if all the repositories of metadata would exploit the same open and standard vocabularies, we have integrated metadata from 4 repositories (DataHub, CLARIN, LRE Map and Metashare) into one RDF repository and strived for a first minimal harmonization of their metadata by mapping the proprietary data models to DCAT and Dublin Code to establish a minimum layer of interoperability at which the data can be uniformly queried and compared.

5.2.1 Targeted resources

For the goal of harmonizing resources we target the following resources which contain metadata about language resources on the web. We will briefly describe the license and format of the data and how we access this data:

- **DataHub**: DataHub.io covers a wide range of linked data resources across a large number of domains. We limit our data extraction to only those that are marked with certain tags ("llod", "linguistics%20lod", "lexicon", "corpus", "thesaurus", "isocat", "linguistic", "linguistics" and "typology") or are members of the "linguistics" groups. The data is provided as RDF using the DCAT vocabulary under a CC-BY-SA 3.0 license.

- **CLARIN Virtual Language Observatory:** These resource collected from the catalogues of CLARIN partners is made available as CMDI XML under a CC-BY 2.0 license
- **LRE-Map:** This resource contains resource descriptions collected at various conferences in the last few years. Data can currently be accessed by scraping the HTML pages which can be retrieved as a single large table. This data is published under an unspecified open license conforming to the Open Definition.
- **Metashare:** This resource is extracted from partners in the MetaShare project. Currently the data is exported from the University Pompeu Fabra node as RDF. It is licensed under the CC-BY-NC-SA 3.0 license.

5.2.2 RDFization of resources

The first step of harmonization is to ensure that all resources are in RDF so that they may be processed by the same tool chain. Currently CLARIN and LRE-Map are not available from the provider as RDF and so must be converted. In the case of LRE-Map a small python script scrapes the HTML tables and generates RDF/XML as required. For CLARIN, we used an XSLT transformation to convert the result from CMDI XML to RDF. Many CMDI documents contain a large amount of data as 'components', which is differently structured for each source resource used in CLARIN, and as such for the moment, with the exception of DCMI metadata, we apply only a generic XML to RDF conversion.

Scripts used in the harmonization are available at http://github.com/liderproject/metadata-harvesting.

5.2.3 Basic harmonization

Each resource is then harmonized by mapping them to the following properties:

Dublin Core

- Title
- Language
- Rights
- Type
- Issued (also on Catalogue Record)
- Creator (also on Catalogue Record)
- Source
- Description

DCAT

- Distribution
- Catalogue Record
- Access URL or Download URL
- Contact Point

RDFS

- See Also

This means that we looked through the existing conversion scripts and tried to ensure that whenever the source properties were present, the appropriate property above was either generated or added. In the case of Datahub this required no extra work and for CLARIN and LRE-Map this process was incorporated into the RDFization of the resource. For Metashare a secondary Python script was used to add appropriate properties.

5.2.4 Further Harmonization

In the next phase of the project, we will strive to improve the harmonization of these resources. In particular we would look to expand the number of properties harmonized to including properties for (linguistic) use and annotation scheme, building on the Metashare ontology currently developed by the LIDER project. Furthermore, we will homogenize the reference to language using unique language codes in place of English names. Furthermore, we will work with content providers to continually obtain the language resources as repositories get updated. Currently, negotiations with Metashare and LDC/ELRA are ongoing to get access to the data.

6 Conclusion and Next Steps

In this deliverable we have presented guidelines for the publication of multilingual data as linked data. We have on the one hand provided guidelines with respect to the following crucial aspects when publishing a multilingual dataset: i) relevant vocabularies to use (grouped in three layers: general vocabularies, linguistic vocabularies and resource-specific vocabularies), ii) best practices for naming, iii) best practices for dereferencing multilingual linked datasets, iv) best practices for encoding textual content, v) best practices for linking and vi) best practices for language identification. All these best practices have been developed as part of a community effort coordinated by the BPMLOD community group, in which people from the LIDER project are playing a leading role (Jorge Gracia from UPM and John McCrae from Bielefeld University).

We have further described our effort to develop a general metamodel that supports the description of metadata for complex datasets. This has in particular resulted in the development of DataID, which builds on existing W3C vocabularies such as PROV-O and VoID. We have further described our efforts in developing a general metamodel for describing linguistic resources in a bottom-up and community driven fashion, starting from the Metashare ontology developed by UPF in Barcelona. We have further made clear recommendations on how to add licensing information to this metadata.

We have also described guidelines for converting legacy linguistic resources (including wordnets, multilingual lexica, multilingual lexical networks as well as terminological resources) into Linked Open Data formats, providing appropriate proof-of-concept implementations for each of these types of resources (Princeton WordNet, Apertium lexica, Babelnet as well as IATE, respectively).
All these guidelines have been developed through involvement of relevant stakeholders and communities through the LD4LT, BPMLOD and ontolex community groups. Industrial stakeholders have been involved in this process.

We have further provided first guidelines for how legacy NLP services can be exposed as LLOD services, taking NIF as input and delivering NIF as output.

Towards developing an ecosystem in which services can automatically discover and extract relevant data from the LLOD, we have first analyzed the status of existing repositories of linguistic metadata, including DataHub, Metashare, LRE Map, Clarin, and LDC, and discussed their pros and cons. We have also described how we have mapped the metadata in four important linguistic repositories (Metashare, CLARIN, LRE Map and DataHub) into RDF by using the DCAT vocabulary as a basic core and least one common denominator for all datasets. This will provide the basis for future efforts to establish interoperability between all these repositories.

As a conclusion, we can issue a number of clear recommendations to providers of linguistic metadata repositories:

- Build on RDF to describe metadata of resources and expose the data through standard Web-compliant interfaces (e.g. SPARQL endpoint, content negotiation).
- Rely on standard W3C vocabularies (e.g. DCAT, Dublin Code) to ensure a minimum level of interoperability between repositories
- Expose rich metadata in terms of licensing (using ODRL), provenance (using PROV-O) and dataset structure (DataID)
- Encourage data providers to release data in well-known, standard and open formats, preferably RDF (but other well-known, standard and open formats such as XML are also acceptable) and to provide a URL as part of the resource metadata under which the resource can be directly downloaded.
- Add rich metadata to characterize the type of resource, its intended use, etc.

For all these aspects, we recommend to follow the best practices and guidelines described as part of this deliverable.

Our next steps include the following:

- Continue our community building efforts and engagement of stakeholders to further refine the guidelines developed so far, making them more robust by considering additional datasets, use cases, etc.
- While in the first phase we have focused on the analysis of current linguistic repositories, the next step will be to focus on discovery and exploitation of LLOD, providing best practices and guidelines for providers of linguistic metadata to describe resources in such a way that they are easily discoverable and services can extract the relevant data to perform some task.
- Continue our effort to disseminate and apply best practices for representing rich metadata using standard W3C vocabularies, in particular for licensing and provenance issues.
- Proceed in the development of a general ontology for linguistic metadata that we have kicked off with the MetaShare ontology developed by UPF in Madrid.
- Extend our efforts to homogenize data from different linguistic metadata repositories by mapping the attributes used in existing repositories to the general linguistic metadata ontology developed as part of the project, clearly going beyond interoperability at the level of DCAT.
- Show the benefits of adopting DataID for the description of complex datasets

- Develop further guidelines for exposing corpora as RDF datasets building on the NLP Interchange Format (NIF)

	Datahub	META-SHARE	LRE Map	CLARIN	LDC
Repository Content					
Domain specific repository	✗	✓	✓	✓	✓
Data Categorization	editable by everyone (groups, tags)	fixed categories according to compiled metadata	fixed categories	fixed categories via ISOcat	fixed categories
Effort to Understand Repository Structure	quick reading up on repository description necessary	intense reading up on repository description necessary	low	high	low
Ease of Dataset Access	easy, open access - search box on main page	easy - 2 steps through the network	easy - search box on main page	difficult - restricted 3 step authorization process	easy - language resources on website navigation
Custom Visualizations (via API)	✓	✗	✗	✓	✗
Data Validation	✗	✗	✗	✗	✗
Accessibility					
User Registration	free - required for data providers only	restricted - required for data providers only	restricted - required for data providers and users	restricted - for countries, institutions only, registration takes 2 days	restricted - universities, foundations, organisations only, with membership fee
Openness of Data	LOD	free and closed licenses	free, redirect to data source	closed data only	closed data only - purchaseable for nonmembers
Dataset Hosting	✓	✓	✗	✓	✓
Data Contribution					
Openness of Dataset Entry	everyone	members only	members only	members only	everyone, but controlled by LDC before upload
Openness of Metadata Entry	everyone	members only	members only	members only	members and providers only
User Voting on Metadata	✗	✗	✗	✗	✗
Bulk Editing	✗	✗	✗	✗	✗
Provenance Constitution					
Kind of Provenance	source URL	information on dataset creation	source URL	information on dataset creation	information on dataset creation, category, source
Derived Data Specification	✗	✗	✗	✗	✗
Provenance Support	rudimentary (source homepage only)	no provenance	rudimentary (source homepage only)	mandatory provenance	fine-grained mandatory provenance
Metadata Processing					
Metadata Maintainance	manually (crowd-sourced by all stakeholders)	manually, only data provider	manually, only LREC community	manually, only CLARIN partners	manually, only LDC members
Automatic Generation of Metadata (link analysis)	✗	✗	✗	✗	✗
Granularity of Metadata	per dataset	1 to 5 per dataset	16 core elements per dataset	15 core elements per dataset	7 elements per dataset

Table 1: Overview of linguistic data repositories

APPENDIX

1 Converting WordNets to Linked Data

Source:
https://www.w3.org/community/bpmlod/wiki/Converting_WordNets_to_Linked_Data
Accessed: 14/10/2014

Converting WordNets to Linked Data

From Best Practices for Multilingual Linked Open Data Community Group

Contents

What is a WordNet?

WordNet is still one of the most widely used lexical resources within natural language processing. From the time since the first version of WordNet was released, many resources have been produced that represent complementary information to WordNet or extend it to other languages .

WordNet is a large lexical database of English nouns, verbs, adjectives and adverbs. Word forms are grouped into more than 117,000 sets of (roughly) synonymous word forms, so called *synsets*. These are interconnected by bidirectional arcs that stand for lexical (word-word) and semantic (synset-synset) relations, including hyper/hyponymy (*tree-oak*), meronymy (*tree-branch*), antonymy (*long-short*) and various entailment relations (*buy-pay, show-see, untie- tie*).

WordNet's synsets and its network structure yield a rough measure of semantic similarity among words and concepts in terms of synset membership as well as the number of arcs separating synsets.

Due to its availability under open licenses, WordNet has become a popular tool for Word Sense Disambiguation (WSD) and Natural Language Processing in general. WordNets have been built for around 100 different languages. Most are mapped onto the Princeton WordNet, enabling translation on the lexical level as well as cross-lingual WSD and applications. WordNet continues to evolve both in terms of coverage and representation of meaning. Recent enhancements include the addition of internet language and partially compositional multi word units. Finally, WordNet has been mapped to formal ontologies, including SUMO and KYOTO .

Selection of vocabularies
lemon

lemon is a model that has been proposed for the representation of lexicons relative to ontologies. As such, this model is well suited to the representation of semantic networks such as WordNet and defines many useful features for linking a WordNet to wider objects in the Semantic Web/Linked Open Data Cloud. *lemon* models lexicons by means of a core consisting of the following elements:

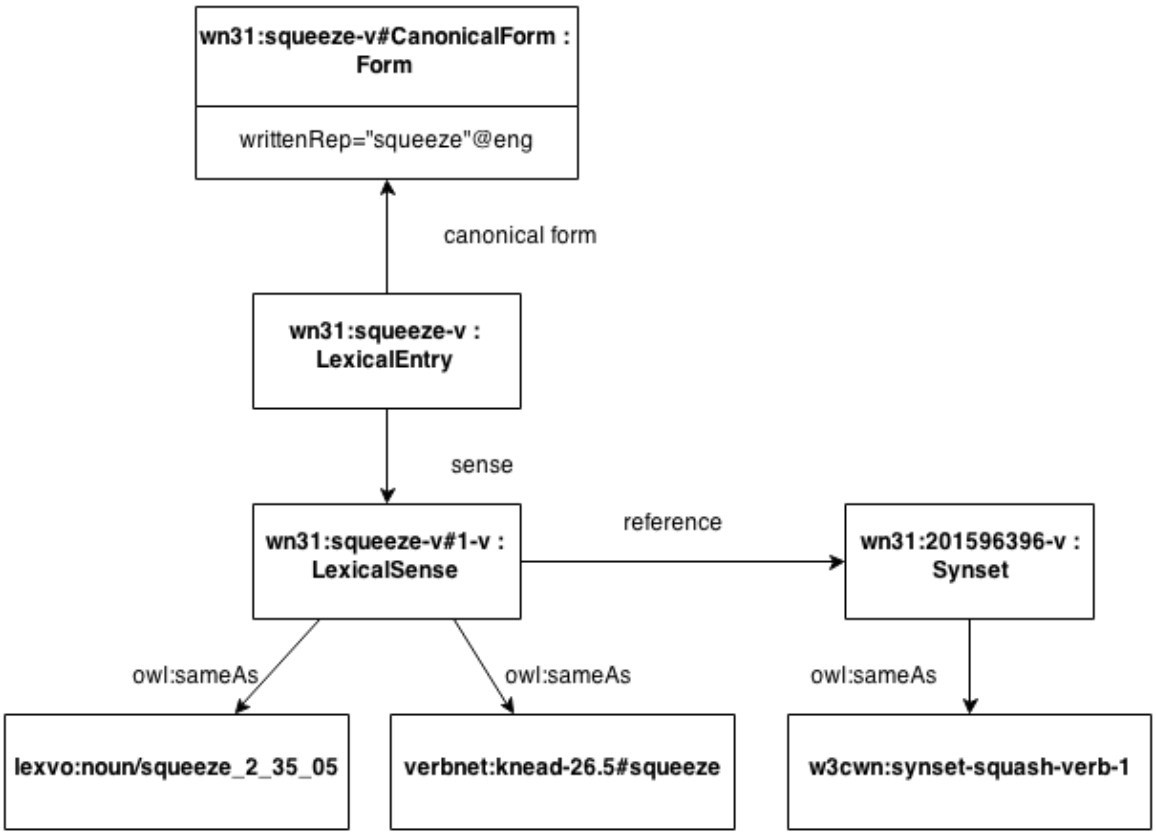
- A *lexical entry* which represents a single word or multi-word unit.
- A *lexical sense*, representing a meaning of that word, which contains a *reference* to a concept in the ontology.
- *Forms*, which are inflected versions of the entry, and associated with a string *representation*.

In fact, in previous work *lemon* has been used not only to represent WordNet but to integrate it with more syntactically sophisticated resources such as VerbNet. As such *lemon* shows potential to help in the integration of lexical data across many levels and languages. The *lemon* model is highly compatible with the ISO standard LMF and forms the basis of the work of the W3C OntoLex Community Group http://www.w3.org/community/ontolex .

SKOS & Custom Vocabulary

In addition, to using the *lemon* vocabulary to model the semantics of WordNet, we use the SKOS vocbulary as this is better suited to WordNet's structural model than a formal ontology language such as OWL. Furthermore, we introduce a new vocabulary at http://wordnet-rdf.princeton.edu/ontology# to include properties found only in WordNet.

RDF Generation
Data modelling



caption An example of the modelling a single word and synset and links to other resources

It is not trivial to apply *lemon* to the case of a WordNet as there is no clear ontology in WordNet. Clearly, WordNet's words can be regarded as *lemon* lexical entries and the word senses correspond well to *lemon*'s lexical senses. WordNet has lemmas and a separate list of variants of these, and as such we create a canonical form for each lemma and a *Form* object for each of these variants. Since there is currently no indication in WordNet of what grammatical properties these variants have, we do not attach additional properties to these variants/forms. As *lemon* is a model for ontology-lexica, the main question is what the reference of the lexical senses should be. We choose to regard WordNet's synsets as ontological references, but instead of assigning them a formal ontological type (e.g., class, property or individual), we introduce a new type Synset as a subclass of *Concept* in SKOS .

This allows us to capture the nature of synsets without ontologizing the semantic network as in . Similarly, we introduce relations such as hypernymy, meronymy etc. as new properties rather than attempt to relate them to existing ontological properties such as OWL's subClassOf. In order to capture the new properties, we introduce an ontology http://wordnet-rdf.princeton.edu/ontology

describing the new properties and classes and provide axioms for the use in the context of both *lemon* and SKOS. These axioms including stating transitivity constraints and equivalence to other vocabularies, e.g., WordNet's *hypernym* to SKOS's *broader*.

```
Furthermore, we link the elements in the ontology to data categories from ISOcat
following the guidelines of .
```

URI design

Another key question concerns the identifiers we use for each element in the data. We do not follow previous exports such as in assigning new identifiers but instead attempt to use the existing identifiers in WordNet. Furthermore, as WordNet has released several versions and is still under development, we consider it important to include the version number in the URI. As such, we use the following scheme for URIs:, as exemplified below:

- Each lexical entry is represented by means of the URL-encoded lemma and then a dash followed by the part-of-speech as a single letter (i.e., 'n(oun)', 'v(erb)', 'a(djective)', 'r (adverb)', 'adjective s(atellite)' or 'p(article)').
- Senses and forms in the model use the entry URI and add a fragment identifier. For forms for which there is no previous identifier in WordNet, we use CanonicalForm and Form-n where n is a number. For senses, the fragment is the index of the senses and the part of speech.
- Synsets are similarly are identified by a number consisting of 8 or 9 digits corresponding to offset codes in the WordNet database The 9 figure codes include an extra initial digit for part-of-speech, followed by a dash and the part of speech as a single letter.

Examples of this scheme include:

```
http://wordnet-rdf.princeton.edu/wn31/cat-n http://wordnet-rdf.princeton.edu/wn31/cat-
n#CanonicalForm http://wordnet-rdf.princeton.edu/wn31/cat-n#2-n http://wordnet-
rdf.princeton.edu/wn31/300001740-a
```

Linking

In addition to providing a RDF/Linked Data version of WordNet, we have incorporated a number of links to other resources. In particular we include the following elements:

- For verbs, we include mappings to VerbNet if they exist. As VerbNet does not currently have a linked data version, we link to the PHP page of the web site.
- We include translations from Open Multilingual WordNet as simple labels on the synsets, identified by the use of language codes.
- We have included previous mappings to LexVo using the current identifiers in WordNet.
- We include links to the W3C WordNet 2.0 export .

- We have created new links to lemonUby .

In addition to these links, we provide support for legacy resources by adding URL mappings from previous versions of WordNet identifiers to the most recent version, with mappings based on. We intend to continue to expand this linksets with contributions from the community.

Publication

The data is made available through the Yuzu http://github.com/jmccrae/yuzu framework, which allows for custom HTML wrapper to be put over a generic linked data site. In this case, this allows the data to be accessible using content negotiation as either HTML with RDFa annotations, RDF/XML, Turtle, N-Triples or JSON-LD formats. In addition, the data is also available as a single zipped N-Triples file. The main database is served from a SQLite database, but the Yuzu framework supports SPARQL querying either over this database or over an external endpoint (the second option is currently in use).

2 Guidelines for Linguistic Linked Data Generation: Multilingual Dictionaries (BabelNet)

Source: http://bpmlod.github.io/report/multilingual-dictionaries/BabelNet/
Accessed: 14/10/2014

Guidelines for Linguistic Linked Data Generation: Multilingual Dictionaries (BabelNet)

Draft Community Group Report 22 September 2014

Editors:

Tiziano Flati, LCL group, Sapienza University of Rome
Roberto Navigli, LCL group, Sapienza University of Rome
Paola Velardi, LCL group, Sapienza University of Rome

Abstract

This document is aimed to guide in the process of creating a Linked Data (LD) version of a lexical resource, in particular BabelNet. These guidelines contain advice on the vocabularies selection, RDF generation process, and publication of the results. As result, the converted language resource is more interoperable and easily accessible on the Web of Data by means of standard Semantic Web technologies. This document describes the models used and the design decisions taken during the conversion of BabelNet into the well-known lemon representation. More in general, we will describe common patterns that naturally emerge when converting a lexical resource into RDF format.

Status of This Document

This specification was published by the Best Practices for Multilingual Linked Open Data. It is not a W3C Standard nor is it on the W3C Standards Track. Please note that under the W3C Community Contributor

This document was published by the Best Practices for Multilingual Linked Open Data community group. It is not a W3C Standard nor is it on the W3C Standards Track.

There are a number of ways that one may participate in the development of this report:

- Mailing list: public-bpmlod@w3.org
- Wiki: Main page
- More information about meetings of the BPMLOD group can be obtained here
- Source code for this document can be found on Github.

If you wish to make comments regarding this document, please send them to http://lists.w3.org/Archives/Public/public-bpmlod/@w3.org (subscribe, archives).

Table of Contents

1. Description of the type of resource

BabelNet is a a very large multilingual encyclopedic dictionary and ontology covering 50 languages, and created by 1) the automatic, seamless integration of WordNet with Wikipedia, OmegaWiki, Open Multilingual WordNet, Wiktionary, and Wikidata and 2) the use of statistical machine translation to acquire a very large amount of multilingual concept lexicalizations.

The backbone model: lemon

We have chosen lemon as the backbone of BabelNet lexical knowledge RDF representation. Lemon is a model proposed for representing lexical information relative to ontologies and for linking lexicons and machine-readable dictionaries to the Semantic Web and the Linked Data cloud. However, we point out that the choice of the models and the definition of properties got refined as the conversion work went ahead.

2. Selection of vocabularies

In the following we list the reference models used during the conversion and provide i) in parenthesis the prefix adopted throughout this document; ii) the URL to the model specification.

Table 1: Namespaces of the vocabularies used along this document

Namespace	prefix	URL
BabelNet-lemon	**bn-lemon**	<http://babelnet.org/model/babelnet#>
Lemon	**lemon**	<http://www.lemon-model.net/lemon#>
SKOS	**skos**	<http://www.w3.org/2004/02/skos/core#>
LexInfo	**lexinfo**	<http://www.lexinfo.net/ontology/2.0/lexinfo#>
Rdf-schema	**rdfs**	<http://www.w3.org/2000/01/rdf-schema#>

Table 1: Namespaces of the vocabularies used along this document

Namespace	prefix	URL
Dublin core	**dc**	<http://purl.org/dc/elements/1.1/>
Dublin terms	**dcterms**	<http://purl.org/dc/terms/#>

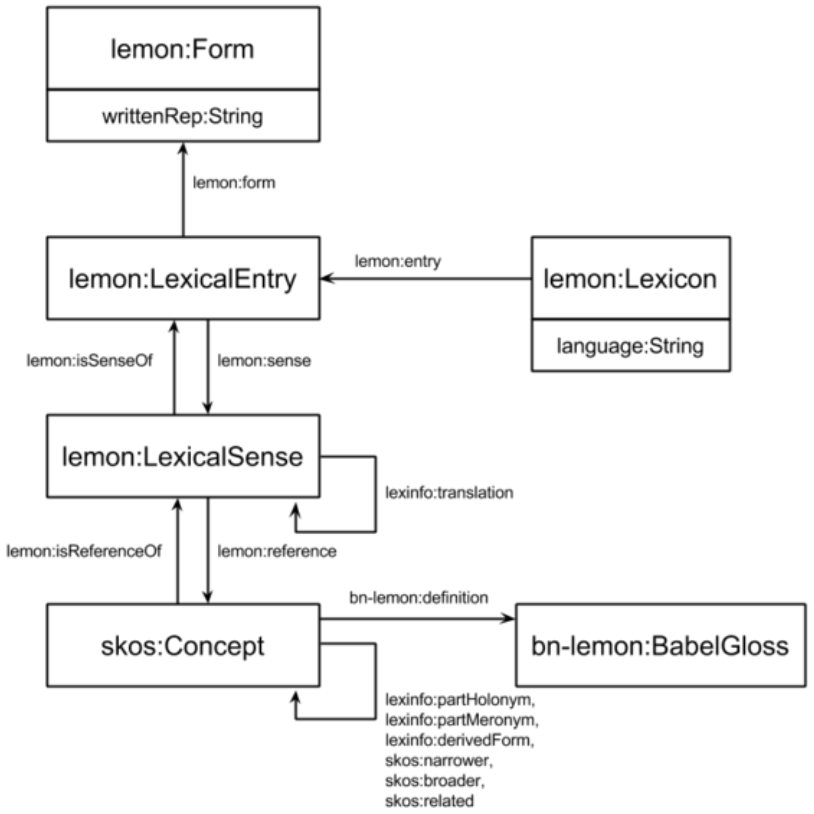
3. Linked Data generation process Technical details

In order to convert BabelNet data into RDF format we need to:

1. **Read the original data**: BabelNet's data, originally stored within Lucene indexes, were accessed through BabelNet's API and translated into RDF triples through the Jena API. The conversion module iterates over the Babel synsets and flushes converted data into data chunks (20k appeared to be a good setting).
2. **Convert the data into RDF**: serialisation format is n-triples (best for huge data sets), and files are printed in compressed format (gz was chosen, since bz2 is not supported by virtuoso) so that the export was compatible with virtuoso loading capacities. The resource is exported into different files according to type of license. The distribution of triples under different licenses is handled via different (Jena) models. In order to attach all BabelNet information to the NC-SA license file, pointers to Babel synsets are moved to the right file before printing (reorganizeModels method in BabelNetLemonFactory class).
3. **Load data into a Virtuoso server**: after the RDF data has been generated, we installed and configured a Virtuoso server and finally loaded the file into the server.

Data modelling and conversion

We first provide a general picture which will help the reader throughout the guidelines and serves as a graphic representation of the main entities and the associated properties involved.



In the following we list the entities and properties chosen for representing the respective pieces of information (words, senses, glosses, etc.), with a brief description and an example. We addressed issues like:

- How do I model my own custom lexicon?
- How do I model word senses and the mapping between a lexical entry in my resource and its senses?
- How do I model common-usage and/or custom relationships between senses?
- How do I insert additional information into the model, such as textual definitions, concept attributes, etc.?

How to model a multilingual lexicon?

Item(s) to model: BabelNet implicitly provides a large multilingual lexicon.

Our solution: The closest entity in the lemon model is **lemon:Lexicon**. The RDF resource consists of a set of Lexicons (lemon:Lexicon), one per language, seen as containers of words. Currently BabelNet supports 50 languages, so BabelNet-lemon includes 50 lemon:Lexicons.

Issues: lemon:Lexicon forces us to work on a language-by-language basis, whereas in BabelNet this distinction does not need to be made explicit, as BabelNet is merely a collection of Babel synsets, i.e. multilingual synsets, and their relations.

How to model word forms and lemmas?

Item(s) to model: BabelNet contains lemmas as elements of its Babel synsets.

Our solution: Lexicons gather Lexical Entries (lemon:LexicalEntry) which comprise the forms of an entry in a certain language (in our case: words of the Babel lexicon). For example the English noun “plane” is encoded as follows:

```
bn:plane_n_EN  a            lemon:LexicalEntry ;

      rdfs:label          "plane"@en ;

      lemon:canonicalForm   <http://babelnet.org/2.0/plane_n_EN/canonicalForm> ;

      lemon:language       "EN" ;

      lemon:sense          <http://babelnet.org/2.0/plane_EN/s00016196n> ,
<http://babelnet.org/2.0/plane_EN/s00062766n> ,

                     <http://babelnet.org/2.0/plane_EN/s00062768n> ,
<http://babelnet.org/2.0/plane_EN/s00062767n> ,

                     <http://babelnet.org/2.0/plane_EN/s00001697n> ;

      lexinfo:partOfSpeech  lexinfo:noun .
```

It could be noted that instead of generating new lexical entries, it could be possible to point to existing entries in some lexical resource (such as Dbnary) which already contains all the information associated with the lexical entry, thus avoiding redundancy of information in BabelNet-lemon.

Lexical Forms (**lemon:Form**), instead, encode the surface realisation(s) of Lexical Entries (in our case: lemmas of Babel words). For instance, the English canonical form "plane" is encoded as:

```
bn:plane_n_EN  lemon:canonicalForm  <http://babelnet.org/2.0/plane_n_EN/canonicalForm> .



<http://babelnet.org/2.0/plane_n_EN/canonicalForm>

      a              lemon:Form ;

      lemon:writtenRep  "plane"@en .
```

Issues: BabelNet does not currently provide all word forms for a lemma, resulting therefore in a duplication of information where each lemon:LexicalEntry (already lemmatized) is associated with its canonical lemon:Form. This is not necessarily an issue, since it is not very clear whether including all the possible morphological forms is truly desirable or not from a lexicographic point of view. Since many languages (e.g., Spanish, Italian and, even worse, Russian or Turkish) do easily spawn tens of different forms for each lemma, the resource would quickly be overwhelmed with too many forms.

How to model a word sense?

Item(s) to model: Babel synsets are sets of word senses expressed in different languages (called Babel senses).

Our solution: Lexical Senses (**lemon:LexicalSense**) represent the usage of a word in a given language as reference to a specific concept (in our case: Babel senses). For instance, the first sense of "plane" in BabelNet is encoded as:

```
<http://babelnet.org/2.0/plane_EN/s00001697n>
    a           lemon:LexicalSense ;
    dc:source       <http://omegawiki.org/> , <http://wordnet.princeton.edu/> ;
    dcterms:license  <http://wordnet.princeton.edu/wordnet/license/> ,
<http://creativecommons.org/licenses/by/3.0/> ;
    lemon:reference  bn:s00001697n .
```

Issues: in order to reduce the amount of redundancy, we decided to merge senses of the same word - i.e., expressing the same concept - in the same language but obtained from different sources (e.g. plane from OmegaWiki and WordNet in the above example). As a result, multiple source and license information is listed for the Lexical Sense.

How to model sense translations?

Item(s) to model: senses which are translations of other senses within a given Babel synset.

Our solution: Senses (modeled as **lemon:LexicalSense**) might also have translations into other senses in other languages. The lemon model alone does not provide a property for expressing this information, so we resorted to the relation lexinfo:translation within the LexInfo ontology model. For example, the fact that the first English sense of "plane" http://babelnet.org/2.0/plane_EN/s00001697n is translated into the French sense http://babelnet.org/2.0/avion_FR/s00001697n is encoded as:

```
<http://babelnet.org/2.0/plane_EN/s00001697n>
    lexinfo:translation <http://babelnet.org/2.0/avion_FR/s00001697n>.
```

LexInfo is an ontology which describes linguistic information and has been used in BabelNet-lemon to represent various linguistic information, such as translation relations and more specific relation types such as meronymy or holonymy.

Issues: A first issue concerns whether including the relation *lexinfo:translation* is essential or not. Within a Babel synset any two senses (in two different languages) are always the translation of each other. For example, if you consider the Babel synset with ID bn:00000356n, the two senses dwelling (in English) and abitazione (in Italian) both belong to the synset and are each the translation of the other; this does not happen for these two senses only, but in general for all the pairs of senses with different languages in the same synset. This fact points out that this information could actually be derived as follows: whenever a system has (i) two*lemon:LexicalSenses* which (ii) belong to the same *skos:Concept* and which (iii) have a

different *lemon:language*, then the system can automatically infer that the two senses are in fact one the translation of the other. This argument undermines thus the necessity of such a translation relation and highlights the possible problem of redundancy. However, having the translation relation could also be a benefit for two reasons: first, because the information is explicit in the resource and no inference would be needed at all; second, because future, subsequent releases of the lexical resource could also refine this relation and, in that case, the specification of the translation relation would be unavoidable.

A second issue concerns the provenance and the confidence information associated with each translation relation. BabelNet's translations come from explicit resource information (e.g., Wikipedia interlanguage links) or from the automatic translations of semantically annotated corpora. We do have a confidence for each of these translations together with the source of the original text. This produces already a distinction regarding the quality and the origin of the translation information. So, despite the resource could potentially include it, the information about translation confidence (was it humanly or automatically produced? by whom? if automatic, with what confidence score?) and translation provenance (what text(s) does the translation come from? who translated and with what tool?) are currently missing.

In addition, translations could be validated through human annotations over time (and thus made more authoritative) and, more in general, the resource could accommodate additional translations coming from different inputs, at different times, from different sources. The general scenario is then that of a set of provisional translations which have different characteristics about quality and provenance. At the moment the translation information is strictly bound to the Babel sense it refers to and models the probability of the sense to belong to a synset. In case of such a general scenario, a best practice is to reify the translation relation into an entity and then attach as many metadata information as needed to the reified relation. The translation entity should in fact model characteristics of the translation process rather than of the target lexical entry itself. In order to account for all such information, the International Tag Set (http://www.w3.org/TR/its20/) stands out as a very good candidate. Thus, to include information about the provenance of a given annotation we could adopt the *its:annotatorsRef* attribute which "provides a way to associate all the annotations of a given data category within the element with information about the processor that generated those data category annotations" (fromhttp://www.w3.org/TR/its20/#provenance). This information should then also be paired with a confidence score (*its:mtConfidence* attribute) certifying the accuracy of the translation that the translation processor (either an automatic tool or a physical person) has provided (see http://www.w3.org/TR/its20/#mtconfidence).

Another possible design choice to represent explicit translations as linked data is to conider using the lemon translation module - currently under development - which "consists essentially of two OWL classes: Translation and TranslationSet. Translation is a reification of the relation between two lemon lexical senses. The idea of using a reified class instead of a property allows us to describe some attributes of the Translation object itself" (from http://www.w3.org/community/ontolex/wiki/Translation_Module, cf. http://lemon-model.net/lemon-cookbook/node18.html).

Future conversions of BabelNet might well include all these additional metadata information with the most suitable model entities.

How to encode concepts?

Item(s) to model: Babel synsets, i.e. sets of multilingual lexicalizations denoting a certain concept, are the core elements of BabelNet.

Our solution: We used SKOS Concepts (skos:Concept) to represent 'units of thought' (in our case: Babel synsets). This was done thanks to its definition and because of its use to model similar objects in other RDF resources (e.g. WordNet). For example, the Babel synset which contains the first sense of plane, i.e.,http://babelnet.org/2.0/s00043466n, is encoded as:

```
bn:s00001697n  a              skos:Concept ;

      bn-lemon:synsetID       "bn:00001697n" ;

      bn-lemon:synsetType     "concept" ;

      dcterms:license         <http://creativecommons.org/licenses/by-nc-sa/3.0/> ;

      lexinfo:partMeronym     bn:s00081337n , bn:s00031553n , bn:s00036743n , bn:s00036922n ,
bn:s00012036n , bn:s00049869n , bn:s00057076n , bn:s00000632n , bn:s00065857n , bn:s00081307n ;

      skos:broader            bn:s00043466n ;

      skos:exactMatch         lemon-Omega:OW_eng_Synset_9672 , dbpedia:Fixed-wing_aircraft , lemon-
WordNet:wn30-02691156-n .
```

Issues: versioning is currently an issue, as we do not have a mechanism to keep track of previous versions of the same synset, if any, and when, i.e. from which version, the synset started to exist in BabelNet.

How to encode concept attributes?

Item(s) to model: associated with a Babel synset, BabelNet has the notion of "concept type", i.e., a type label which declares the concept either as a 'Concept' (e.g., "singer") or a 'Named Entity' (e.g., "Frank Sinatra").

Our solution: to this end we provided a new property in our own BabelNet-specific RDF vocabulary, called **bn-lemon:synsetType**. In the above example, the fact that the previous synset represents a concept is encoded by:

```
      bn-lemon:synsetType     "concept" ;
```

Issues: since we could not find any similar notion in the models used, we decided to introduce a new property. In general, since attributes can bear arbitrary information which might or might not fit pre-defined entities and properties, it is not possible to give a general guideline in this case and it is thus responsibility of the designer to find the best solution, on a case-by-case basis.

How to model a concept gloss?

Item(s) to model: BabelNet provides multiple glosses in several languages for each Babel synset. A gloss is a short explanatory sentence of a concept. For example the English Wikipedia definition for the first sense of plane in BabelNet is "A fixed wing aircraft is an aircraft capable of flight using wings that generate lift due to the vehicle's forward airspeed and the shape of the wings".

Our solution: we defined a new entity, called **bn-lemon:BabelGloss**, which encodes a textual definition associated to a Babel synset. The property bn-lemon:definition binds synsets to their gloss(es). The fragment of text below is intended to show an example of the encoding of an English BabelGloss. Note that information such as the reference language (lemon:language) and the source of the definition (dc:source) are also attached to the gloss.

```
bn:s00001697n_Gloss3_EN

      a              bn-lemon:BabelGloss ;

      bn-lemon:gloss   "A fixed wing aircraft is an aircraft capable of flight using wings that generate lift due to the vehicle's forward airspeed and the shape of the wings." ;

      dc:source       <http://wikipedia.org/> ;

      dcterms:license  <http://creativecommons.org/licenses/by-sa/3.0/> ;

      lemon:language   "EN" .


bn:s00001697n  bn-lemon:definition  bn:s00001697n_Gloss3_EN .
```

Issues: since there might well be more than one gloss in a certain language for a given Babel synset (coming from different sources, such as Wikipedia or OmegaWiki), bn-lemon:BabelGloss's URIs include an incremental integer. Another choice would have been to include a source identifier ('Wiki', 'Omega', 'WordNet', etc.) within the gloss's URI (such as, for instance, bn:s00001697n_Gloss_Wiki_EN or bn:s00001697n_Gloss_Omega_EN).

How to model semantic relations?

Item(s) to model: BabelNet comes with a very high number of semantic relations, also characterized by their semantic type. Relation types are basically inherited from WordNet and include, among others, hypernymy (is-a), hyponymy (has-a), meronymy (is-part-of), holonymy (has-part) and even derivationally related forms (such as 'solve#v' for 'solution#n'). Most of the edges, though, lack a clear typing and are labelled as mere "related-to" edges.

Our solution: In order to describe the several types of semantic relations that a synset is involved in, we exploited both the LexInfo and the SKOS models. In fact, relations such as meronymy, holonymy and derivationally related forms can be found in the LexInfo model (**lexinfo:partMeronym**, **lexinfo:partHolonym** and**lexinfo:derivedForm**, respectively), while all the other types, such as hypernymy, hyponymy, and the more general un-typed relatedness, have been drawn from the SKOS model (**skos:narrower**, **skos:broader** and **skos:related**, respectively). As regards the above example, we show an excerpt encoding several semantic relation types:

```
 lexinfo:partMeronym      bn:s00081337n , bn:s00031553n , bn:s00036743n , bn:s00036922n , bn:s00012036n , bn:s00049869n , bn:s00057076n , bn:s00000632n , bn:s00065857n , bn:s00081307n ;

      skos:broader       bn:s00043466n ;

      skos:exactMatch    lemon-Omega:OW_eng_Synset_9672 , dbpedia:Fixed-wing_aircraft ,
```

Issues: we also note that there is another type of relation encoding the notion of 'equivalence' between concepts across different resources (such as the BabelNet synset "bn:00001697n" and the DBpedia concept http://dbpedia.org/page/Fixed-wing_aircraft). We thus decided to describe this notion of equivalence by means of the *skos:exactMatch* property; note, however, that a similar choice could have been made in favor of the *owl:sameAs* property or by relaxing the type of matching with *skos:closeMatch*, *rdf:seeAlso*, etc.

How to encode resource names?

Item(s) to model: Resource identifiers can be encoded by using either URIs or IRIs, strings which uniquely identify resources in a model. URIs facilitate automatic elaboration of linked data, whereas IRIs improve readability for human end users. URIs can either be descriptive, that is, encoding as much meaning as possible (e.g., bn:Haus_n_DE which represents the German lexical entry for "House"), or opaque, that is providing encoding names which do not convey the content of the resource identifier (e.g., the URI for synset with ID bn:00024498n is bn:s00024498n which does not truly say much about the synset's content). On the other hand, IRIs preserve a language's specific alphabet but at the same time hinder readability to non-native speakers. For example the following IRI "bn:楼宇_ZH/s00044994n" encodes the sense of House in Chinese, but a non native speaker can have a hard time understanding this. An additional dimension of the naming scheme is represented by the choice whether to include the language tag in the resource identifier in the path or URI (as in the example above) or in the host name (e.g., http://zh.babelnet.org/2.0/楼宇/s00044994n).

Our solution: use both URIs and IRIs in order to have the highest degree of flexibility and expressivity. Since our lexical resource is not divided up into different datasets, the option to provide the language identifier as part of the host name was not practical; so we decided to include it as part of the URI/IRI. For instance, the previous IRI encoded the language by means of the 'ZH' suffix, concatenated with the sense string 楼宇.

Issues: The current usage of resource identifiers is not unified yet, so that certain entities, such as Babel synsets and BabelGlosses, are encoded with URI, while other language-specific entities, such as BabelSenses and LexicalEntries, use IRI. Generally speaking, whenever it was possible to do so, we preferred meaningful URIs (e.g., bn:Haus_n_DE); in other cases we came up with IDs which uniquely identified the resource. As regards Babel synset URIs, we preferred to mantain the synset identity quite general and avoided to promote any sense as the main sense for that synset. As regards glosses, the gloss URI encodes the synsetID the gloss refers to, the language and an incremental integer which differentiates between glosses of the same language (e.g., the synset for the first sense of "home" has 3 English glosses coming from different sources, identified by bn: s00044994n_Gloss1_EN, bn: s00044994n_Gloss2_EN, and bn: s00044994n_Gloss3_EN).

4. Linked Data Publication

In addition to the data itself, most of the BabelNet data (skos:Concept, lemon:LexicalSense, etc.) has some useful metadata attached to it.

Resource metadata

These are metadata concerning the resource itself and include, for example, the type of license, the release version, the date of the release, the creator authority and the type of resource. What follows shows how this information is embedded:

```
<dcterms:license rdf:resource="http://creativecommons.org/licenses/by-nc-sa/3.0/"/>
<rdfs:label xml:lang="EN">BabelNet</rdfs:label>
<owl:versionInfo>2.0</owl:versionInfo>
<dc:date>October 2013</dc:date>
<dc:creator>Linguistic Computing Laboratory - Computer Science Department - Sapienza University of Rome</dc:creator>
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Ontology"/>
```

License

The dc:license entity provides information about the license information by referring to the "legal document giving official permission to do something with the resource". Values include "http://creativecommons.org/licenses/by-sa/3.0/", "http://wordnet.princeton.edu/wordnet/license/", etc.

Provenance

As stated in the Dublin Core vocabulary, the dc:source contains information about "a related resource from which the described resource is derived". In BabelNet this can take on several values, ranging among "http://omegawiki.org/", "http://wordnet.princeton.edu/", "http://wikipedia.org/", etc.

It is worth noting that also the PROV Ontology (PROV-O) could be adopted for describing provenance metadata, since "The goal of PROV is to enable the wide publication and interchange of provenance on the Web and other information systems". The PROV-O provides in fact "a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts", including versioning information, activities, agents, roles and location identifiers, among others. Thanks to the higher expressivity of PROV-O, one could think to use the model alone for encoding all the information about provenance and licensing. Even if some terms in the Dublin Core vocabulary can be mapped in a one-to-one correspondence to terms in the PROV-O (e.g., *dc:provenance* can be mapped to the PROV-O term *prov:has_provenance*, see http://www.w3.org/TR/2013/WD-prov-dc-20130312/), this is unfortunately not always the case (e.g., *dc:license* has no direct corresponding term in the PROV-O). In addition to this, BabelNet is not fully exploiting the DC's vocabulary expressive power. There are in fact a lot of provenance-related Dublin terms - about who affected a resource (*dcterms:contributor*, *dcterms:publisher*, etc.), about when (*dcterms:created*, *dcterms:modified*, *dcterms:valid*, etc.) and how (*dcterms:license*, *dcterms:rights*, *dcterms:isVersionOf*, etc.) - which are not currently included in our encoding and which instead are worth exploring and exploiting.

In conclusion, if on the one hand PROV-O provides a more complete set of tools for expressing information about provenance, on the other hand it still lacks some aspect concerning licensing and this shows how, eventually, a combination of the two is probably needed.

Versioning

As a final remark, versioning has been left out from the conversion, for the moment. As a first step the entity *owl:versionInfo* could be used so as to provide a textual reference for the current version of the linked data. This entity is currently used only in BabelNet-lemon schema description and globally provides a version number for the whole release (for RDF, currently 2.0). In the real world, though, maybe a more sophisticated infrastructure would be needed in order to express more complex versioning description needs (for example, what should be considered to be different versions of a resource?): a long-standing and

notable example of such a phenomenon is represented by WordNet, where concepts have been split, lumped, deleted or added throughout time across versions. The current available vocabularies, in fact, do not account for heavy changes in the resource and this aspect might thus be investigated in more detail in the next future by the whole community.

5. Data querying

In order to grasp the real power of the resource, we will now introduce some concrete SPARQL queries. Despite their simplicity, the following queries model very common patterns in the industrial panorama. These can then be extended and customized to your specific needs with very little effort.

Retrieve the senses of a given lemma

Given a word, e.g. home, retrieve all its senses and corresponding synsets in all supported languages:

```
SELECT DISTINCT ?sense ?synset WHERE {
  ?entries a lemon:LexicalEntry .
  ?entries lemon:sense ?sense .
  ?sense lemon:reference ?synset .
  ?entries rdfs:label ?term .
  FILTER (STR(?term)="home")
} LIMIT 10
```

Retrieve the senses of a lemma for a certain language

Given a word, e.g. home, retrieve all its senses and corresponding synsets in English:

```
SELECT DISTINCT ?sense ?synset WHERE {
  ?entries a lemon:LexicalEntry .
  ?entries lemon:language "EN" .
  ?entries lemon:sense ?sense .
  ?sense lemon:reference ?synset .
  ?entries rdfs:label ?term .
  FILTER (STR(?term)="home")
} LIMIT 10
```

Retrieve the translations of a given sense

Given a sense, we want to obtain all its translations: e.g., given the sense http://babelnet.org/2.0/home_EN/s00044488n:

```
SELECT ?translation WHERE {
  ?entry a lemon:LexicalSense .
  ?entry lexinfo:translation ?translation .
  FILTER (STR(?entry)="http://babelnet.org/2.0/home_EN/s00044488n")
}
```

Retrieve license information about a sense

For instance, given the sense http://babelnet.org/2.0/home_EN/s00044488n:

```
SELECT ?license WHERE {
  ?entry a lemon:LexicalSense .
  ?entry dcterms:license ?license .
  FILTER (STR(?entry)="http://babelnet.org/2.0/home_EN/s00044488n")
}
```

Retrieve the resources to which sense information belong

For instance, given the sense: http://babelnet.org/2.0/home_EN/s00044488n:

```
SELECT ?source WHERE {
  ?entry a lemon:LexicalSense .
  ?entry dc:source ?source .
  FILTER (STR(?entry)="http://babelnet.org/2.0/home_EN/s00044488n")
}
```

Retrieve textual definitions in all languages

For instance, given the synset: http://babelnet.org/2.0/s00000356n:

```
SELECT DISTINCT ?language ?gloss ?license ?sourceurl WHERE {
```

```
    ?url a skos:Concept .
    ?url bn-lemon:synsetID ?synsetID .
    OPTIONAL {
        ?url bn-lemon:definition ?definition .
        ?definition lemon:language ?language .
        ?definition bn-lemon:gloss ?gloss .
        ?definition dcterms:license ?license .
        ?definition dc:source ?sourceurl .
    }
    FILTER (STR(?url)="http://babelnet.org/2.0/s00000356n")
}
```

Retrieve textual definitions in a certain language

For instance, given the synset: http://babelnet.org/2.0/s00000356n:

```
SELECT DISTINCT ?gloss ?license ?sourceurl WHERE {
    ?url a skos:Concept .
    ?url bn-lemon:synsetID ?synsetID .
    OPTIONAL {
        ?url bn-lemon:definition ?definition .
        ?definition lemon:language "EN" .
        ?definition bn-lemon:gloss ?gloss .
        ?definition dcterms:license ?license .
        ?definition dc:source ?sourceurl .
    }
    FILTER (STR(?url)="http://babelnet.org/2.0/s00000356n")
}
```

Retrieve a synset's hyponyms

For instance, given the synset: http://babelnet.org/2.0/s00000356n:

```
SELECT ?narrower WHERE {
  ?entry a skos:Concept .
  OPTIONAL { ?entry skos:narrower ?narrower }
  FILTER
  (STR(?entry)="http://babelnet.org/2.0/s00000356n")
}
```

Retrieve a synset's hypernyms

For instance, given the synset: http://babelnet.org/2.0/s00000356n:

```
SELECT ?broader WHERE {
  ?entry a skos:Concept .
  OPTIONAL { ?entry skos:broader ?broader }
  FILTER
  (STR(?entry)="http://babelnet.org/2.0/s00000356n")
}
```

Retrieve all the RDF information of a synset

For instance, given the synset http://babelnet.org/2.0/s00000356n:

```
DESCRIBE <http://babelnet.org/2.0/s00000356n>
```

References

[BabelNet at LREC 2014]
M. Ehrmann, F. Cecconi, D. Vannella, J. McCrae, P. Cimiano, R. Navigli, *Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0.*. Proc. of the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26-31 May, 2014.

[BABELNET_SPARQL_ENDPOINT]
BabelNet's SPARQL endpoint.
URL: http://babelnet.org:8084/sparql/

[BABELNET_HOMEPAGE]
BabelNet homepage. URL: http://babelnet.org/

[LEMON_MODEL]
The lemon model. URL: http://lemon-model.net/
[JENA_API]
Jena API. URL: https://jena.apache.org/
[ITS2.0]
International Tag Set. URL: http://www.w3.org/TR/its20/
[PROV_ONTOLOGY]
PROV Ontology. URL: http://www.w3.org/TR/prov-o/

ReSpec2

3 Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries

Source: http://bpmlod.github.io/report/bilingual-dictionaries/index.html
Accessed: 14/10/2014

Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries
Draft Community Group Report 22 September 2014
Editor:

Jorge Gracia, Ontology Engineering Group, Universidad Politécnica de Madrid

Abstract

This document is aimed to guide in the process of creating a linked data (LD) version of a lexical resource, particularly a bilingual dictionary. It contains advice on the vocabularies selection, RDF generation process, and publication of the results. As result of publishing the data as LD, the converted language resource will be more interoperable and easily accessible on the Web of Data by means of standard Semantic Web technologies. The process described in this document has been illustrated with real examples extracted from Apertium RDF, an open-source machine translation system which has their data available for download.

Status of This Document

This specification was published by the Best Practices for Multilingual Linked Open Data. It is not a W3C Standard nor is it on the W3C Standards Track. Please note that under the W3C Community Contributor License Agreement (CLA) there is a limited opt-out and other conditions apply. Learn more about W3C Community and Business Groups.

This document was published by the Best Practices for Multilingual Linked Open Data community group. It is not a W3C Standard nor is it on the W3C Standards Track.

There are a number of ways that one may participate in the development of this report:

- Mailing list: public-bpmlod@w3.org
- Wiki: Main page
- More information about meetings of the BPMLOD group can be obtained here
- Source code for this document can be found on Github.

If you wish to make comments regarding this document, please send them to http://lists.w3.org/Archives/Public/public-bpmlod/@w3.org (subscribe, archives).

Table of Contents

1. Description of the type of resource

The type of language resources covered in this document is *bilingual electronic dictionaries*. A bilingual dictionary is a specialized dictionary used to translate words or phrases from one language to another. They can be unidirectional or bidirectional, allowing translation, in the latter case, to and from both languages. In addition to the translation, a bilingual dictionary usually indicates the part of speech, gender, verb type, declination model and other grammatical properties to help a non-native speaker use the word.

We are interested in bilingual dictionaries that have their data in a machine-processable format, no matter whether it is stored locally or is accessible on the Web (e.g., for download). We assume that the data is represented in a structured or semi-structured way (e.g., relational database, xml, csv, etc.).

We will illustrate our discussion with real examples from the conversion of the Apertium dictionaries into RDF [*AP_RDF*]. Apertium [*AP_PAPER*] is a free/open-source machine translation platform originally designed to translate between closely related languages, although it has recently been expanded to treat more divergent language pairs. There exist Lexical Markup Framework [*LMF*] versions of their linguistic data which can be found here and have been used as starting point for the RDF version.

2. Selection of vocabularies

- We propose ***lemon*** (LExicon Model for ONtologies) [*LEMON*, *LEMON_PAPER*] to model the RDF representation of the linguistic descriptions contained in the bilingual dictionaries. *lemon* has been designed to extend the lexical layer of ontologies with as much linguistic information as needed, and to provide it as linked data on the Web. From *lemon* we take mechanisms to represent *lexicons, lexical entries, forms,* and *lexical senses*.
- The use of *lemon* is complemented with ***Lexinfo*** [*LEXINFO*]. Lexinfo is an ontology of types, values and properties to be used with the lemon model, partially derived from ISOcat. We use Lexinfo as a catalog of data categories (e.g., to denote gender, number, part of speech, etc.).
- We will use the lemon ***Translation Module*** [*TR*, *TR_PAPER*] for representing translations. The translation module consists essentially of two classes: `Translation` and `TranslationSet`. Translation is a reification of the relation between two lemon lexical senses associated to terms in different languages. The idea of using a reified class allows us to describe some attributes of the

Translation object itself, basically: translationSource, translationTarget, translationConfidence, context, and translationCategory.

- ***Translation categories*** are represented by pointing to an external catalog (e.g. to state that a translation is a "cultural equivalent"). We propose the one at [*TRCAT*] but any other could be used instead.
- Other extendedly used vocabularies such as ***Dublin Core*** [*DC*] are used to attach valuable information about provenance, authoring, versioning, or licensing.
- Finally, the ***Data Catalogue Vocabulary*** [*DCAT*] will be used to represent other metadata information associated to the publication of the RDF dataset.

NOTE

Both lemon and the Translation Module are currently under revision by the W3C Ontolex community group [*ONTOLEX*]. Nevertheless, the resultant model is expected to be backwards compatible with the current ones. Thus, the content of this guideline should remain valid for its use with the future model.

We summarize in the following table a list of relevant namespaces that will be used in the rest of this document.

Table 1: Namespaces of the relevant vocabularies

owl	<http://www.w3.org/2002/07/owl#>
rdfs	<http://www.w3.org/2000/01/rdf-schema#>
lemon	<http://www.lemon-model.net/lemon#>
lexinfo	<http://www.lexinfo.net/ontology/2.0/lexinfo#>
tr	<http://purl.org/net/translation#>
trcat	<http://purl.org/net/translation-categories#>
dc	<http://purl.org/dc/elements/1.1/>
dct	<http://purl.org/dc/terms/>
dcat	<http://www.w3.org/ns/dcat#>
apertium	<http://linguistic.linkeddata.es/id/apertium/>

3. RDF generation

For the generation and publication processes we have followed the recommendations in [*GUIDE_MLD*], adapted to our particular case.

3.1 Analysis of the data sources

The first activity of the publication of Linked Data is to analyse and specify the resources that will be used as source of data, as well as the data model(s) used within such sources. The analysis covers two aspects

- *Data model*. All available information about the data model used in the sources has to be analysed, comprising standards, terminologies, etc.
- *Content*. The data underlying such models has to be analysed also, and their linguistic features examined: e.g., to identify language dependent/independent information, to understand how names and identifiers have been constructed in the source data, how language have been encoded, etc.

The result of this phase is strongly dependent on the particular data source and its representation formalism. The general advise would be to get a good understanding of how the original dictionary is represented in order to define proper conversion rules of the original data into RDF.

Regarding our illustrating example (the Apertium EN-ES dictionary), the model used for representing the data is [*LMF*]. The following lines of code illustrate how the content is represented in LMF/XML for a single translation:

EXAMPLE 1: A single EN-ES translation in LMF/XML

```
<Lexicon>    <feat att="language" val="en"/>    ...
<LexicalEntry id="bench-n-en">       <feat att="partOfSpeech"
val="n"/>    <Lemma>           <feat att="writtenForm"
val="bench"/>        </Lemma>       <Sense id="bench_banco-n-1"/>
</LexicalEntry>    ... </Lexicon> <Lexicon>     <feat
att="language" val="es"/>     ...     <LexicalEntry id="banco-n-
es">                  <feat att="partOfSpeech" val="n"/>    <Lemma>
          <feat att="writtenForm" val="banco"/>     </Lemma>
<Sense id="banco_bench-n-r"/>     </LexicalEntry>     ...
</Lexicon> ... <SenseAxis id="bench_banco-n-banco_bench-n"
senses="bench_banco-n-1 banco_bench-n-r"/> ...
```

3.2 Modelling

The first step in the modelling phase is the selection of the domain vocabularies to be used. This has been already discussed in the above section "selection of vocabularies". Next, it has to be decided how the representation scheme of the source data has to be mapped into the new model. In the case of bilingual dictionaries, each dictionary is converted into three different objects in RDF (no matter if the original data comes in one or several files):

- Source lexicon
- Target lexicon
- Translation Set

This is illustrated in the following figure, for the conversion of an English-Spanish bilingual dictionary.

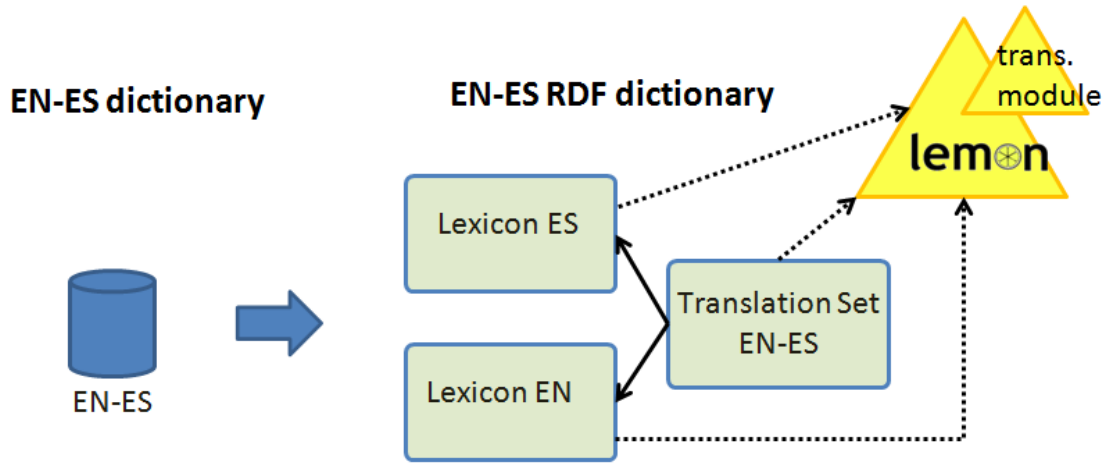


Fig. 1 Conversion of a bilingual electronic dictionary into RDF

In our opinion, this is the division that fits more naturally in the scheme of *lemon* and the translation module. As result, two independent monolingual lexicons will be published on the Web of Data, along with a set of translations that connects them. The publication of additional bilingual dictionaries (following the same scheme) would imply the creation of a pool of online monolingual lexicons that grows with time, all of them potentially connected within the same RDF graph by sets of translations.

Going into the details of the model, the following figure illustrates the representation scheme used for a single translation, in terms of *lemon* and the translation module:

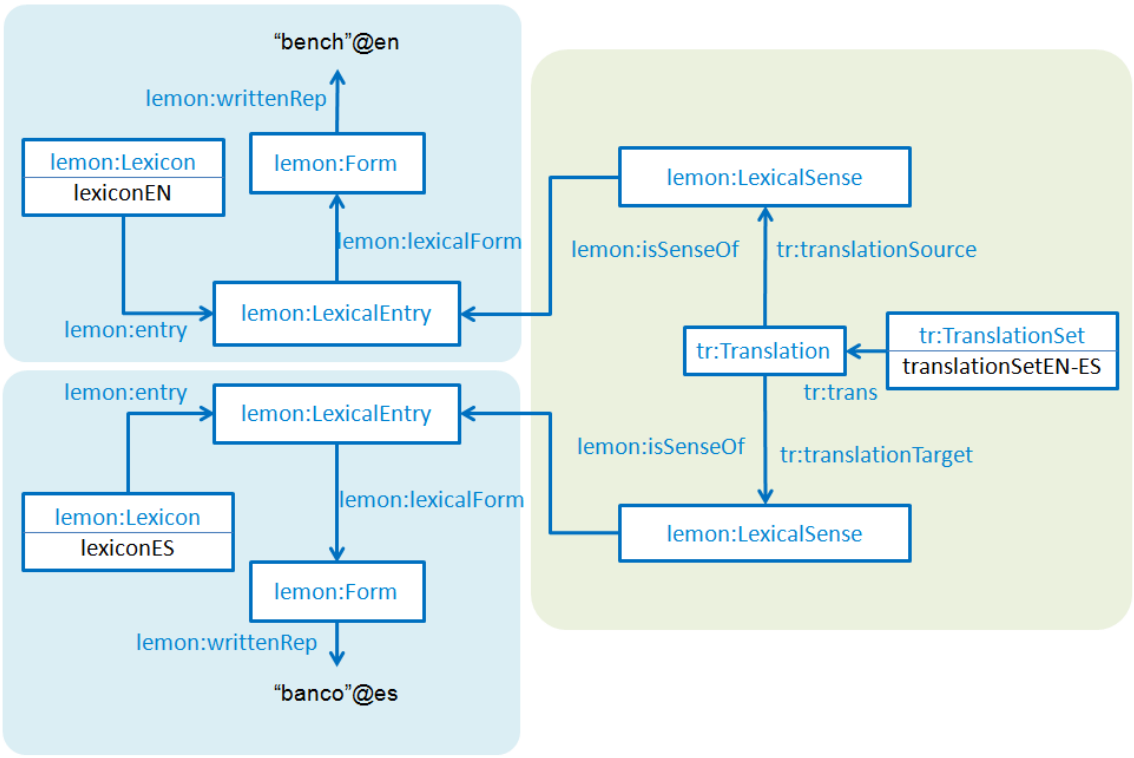


Fig. 2 Modelling a translation in RDF

In short, `lemon:LexicalEntry` and their associated properties are used to account for the lexical information, while the `tr:Translation` class puts them in connection through `lemon:LexicalSense`. Other options are possible, of course, such as connecting directly the lexical entries without defining "intermediate" senses. Nevertheless, we understand that translations occur between specific meanings of the words and the class `lemon:LexicalSense` allows us to represent this fact explicitly.

3.3 URIs design

Among the different patterns and recommendations for defining URIs we propose the one at [*ISA_URIS*] although others could be used instead. In short, the ISA pattern is as follows: `http://{domain}/{type}/{concept}/{reference}`, where `{type}` should be one of a small number of possible values that declare the type of resource that is being identified. Typical examples include: 'id' or 'item' for real world objects; 'doc' for documents that describe those objects; 'def' for concepts; 'set' for datasets; or a string specific to the context, such as 'authority' or 'dcterms'.

In our example, the main components (lexicons and translation set) of the RDF bilingual dictionary are named as follows:

EXAMPLE 2

```
Apertium English lexicon:
http://linguistic.linkeddata.es/id/apertium/lexiconEN  Apertium
Spanish lexicon:
http://linguistic.linkeddata.es/id/apertium/lexiconES  Apertium
```

```
English-Spanish translation set:
http://linguistic.linkeddata.es/id/apertium/tranSetEN-ES
```

In order to construct the URIs of the lexical entries, senses, and rest of lexical elements, we have *preserved the identifiers of the original data whenever possible*, propagating them into the RDF representation. Some minor changes have been introduced, though. For instance, in the original data the identifier of the lexical entries ended with the particle "-l" or "-r" depending on their role as "source" or "target" in the translation. In our case, the directionality is not preserved at the level of Lexicon (but in the Translation class) so these particles are removed form the name. In addition, some other suffixes have been added for readability (this step is optional): "-form" for lexical forms, "-sense" for lexical senses, and "-trans" or translation. See the following section "generation" for particular examples.

3.4 Generation

This activity deals with the transformation into RDF of the selected data sources using the representation scheme chosen in the modelling activity. Technically speaking, there are a number of tools that can be used to assist the developer in this task (see here for a survey), depending on the format of the data source. In our case, Open Refine (with its RDF extension) was used for defining the transformations from XML into RDF.

As result of the transformation, three RDF files were generated, one per component (lexicons and translation set). The following examples contain the RDF code (in turtle) of a single translation. The three pieces of code come from the EN and ES lexicons and from the EN_ES translation set, respectively, of the Apertium example:

EXAMPLE 3

```
apertium:lexiconEN a lemon:Lexicon ;      dc:source
<http://hdl.handle.net/10230/17110> . ... apertium:lexiconEN
lemon:entry apertium:lexiconEN/bench-n-en .
apertium:lexiconEN/bench-n-en a lemon:LexicalEntry ;
    lemon:lexicalForm apertium:lexiconEN/bench-n-en-form ;
    lexinfo:partOfSpeech lexinfo:noun . apertium:lexiconEN/bench-n-
en-form a lemon:Form ;     lemon:writtenRep "bench"@en .
```

EXAMPLE 4

```
apertium:lexiconES a lemon:Lexicon ;      dc:source
<http://hdl.handle.net/10230/17110> . ... apertium:lexiconES
lemon:entry apertium:lexiconES/banco-n-es .
apertium:lexiconES/banco-n-es a lemon:LexicalEntry ;
    lemon:lexicalForm apertium:lexiconES/banco-n-es-form ;
    lexinfo:partOfSpeech lexinfo:noun . apertium:lexiconES/banco-n-
es-form a lemon:Form ;     lemon:writtenRep "banco"@es .
```

EXAMPLE 5

```
apertium:tranSetEN-ES a tr:TranslationSet ;       dc:source
<http://hdl.handle.net/10230/17110> ; ...  apertium:tranSetEN-ES
tr:trans apertium:tranSetEN-ES/bench_banco-n-en-sense-banco_bench-
n-es-sense-trans . apertium:tranSetEN-ES/bench_banco-n-en-sense a
lemon:LexicalSense ;       lemon:isSenseOf apertium:lexiconEN/bench-
n-en . apertium:tranSetEN-ES/banco_bench-n-es-sense a
lemon:LexicalSense ;       lemon:isSenseOf apertium:lexiconES/banco-
n-es . apertium:tranSetEN-ES/bench_banco-n-en-sense-banco_bench-n-
es-sense-trans a tr:Translation ; tr:translationSource
apertium:tranSetEN-ES/bench_banco-n-en-sense ;
    tr:translationTarget apertium:tranSetEN-ES/banco_bench-n-es-
sense .
```

Reproducibility is an important feature, so the mappings between the original data and the new RDF-based model, as well as the scripts for the RDF generation, should be recorded and stored to enable their later reuse.

4. Publication

The publication step involves: (1) dataset publication, (2) metadata publication, and (3) enabling effective discovery. Here we will focus on the second task. In the context of Linked Data, there are two major vocabularies for publishing metadata for describing datasets and catalogues: VoID (Vocabulary of Interlinked Datasets) [*VOID*], and DCAT (Data Catalogue Vocabulary) [*DCAT*]. In principle, we think that DCAT suffices for the purposes of describing the elements generated in the RDF conversion of bilingual dictionaries. Further, some data management platforms such as Datahub use DCAT in a preferred way for representing metadata. In any case, DCAT can be complemented with VoID or other vocabularies if required.

The RDF version of Apertium EN-ES was published in Datahub. The Datahub platform created a metadata file for the Apertium EN-ES dataset based on DCAT. We extended such metadata file with some additional missing information such as provenance, license, and related resources. The extended metadata was publishedas part of the Apertium EN-ES Datahub entry. The following lines are a fragment of it:

EXAMPLE 6

```
<dcat:Dataset
rdf:about="http://linguistic.linkeddata.es/set/apertium/EN-ES">
<owl:sameAs rdf:resource="http://datahub.io/dataset/apertium-en-
es"></owl:sameAs>    <dct:source
rdf:resource="http://hdl.handle.net/10230/17110"></dct:source>
<dct:license rdf:resource="http://purl.oclc.org/NET/rdflicense/gpl-
3.0"></dct:license>    <rdfs:seeAlso
rdf:resource="http://dbpedia.org/resource/Apertium"></rdfs:seeAlso>
<rdfs:seeAlso rdf:resource="http://purl.org/ms-lod/UPF-
MetadataRecords.ttl#Apertium-en-es_resource-5v2"></rdfs:seeAlso>
</dcat:Dataset>
```

5. Data maintenance

TO COMPLETE

6. Recommendations

- Separate the monolingual lexicons from the translation sets (different graphs and/or files).
- Lexical senses should play the role of connectors between translations and lexical entries.
- Be consistent with the rules for naming and URIs creation.
- Keep the identifiers of the legacy data if possible, but removing indicators of directionality if any (e.g., "l", "r", "left", "right", ...)
- ...

References

[AP_PAPER]

M. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. Tyers, *Apertium: a free/open-source platform for rule-based machine translation* . Machine Translation, vol. 25, no. 2, pp. 127-144, 2011.

[AP_RDF]

RDF version of the Apertium bilingual dictionaries.
URL: http://linguistic.linkeddata.es/apertium/

[DC]

DCMI Metadata Terms. URL: http://purl.org/dc/elements/1.1/

[DCAT]

F. Maali, J. Erickson (Eds.). Data Catalog Vocabulary (DCAT). W3C Recommendation. January 2014
URL: http://www.w3.org/TR/vocab-dcat/

[GUIDE_MLD]

A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, and G. Aguado-de Cea, *Guidelines for multilingual linked data* , in Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS'13). New York, NY, USA: ACM, Jun. 2013.

[ISA_URIS]

P. Archer, S. Goedertier, and N. Loutas, *Study on persistent URIs* Tech. Rep., ISA, Dec. 2012.

[LEMON_PAPER]

J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner, *Interchanging lexical resources on the Semantic Web* . Language Resources and Evaluation, vol. 46, 2012.

[LEMON]

The lemon model. URL: http://lemon-model.net/

[LEXINFO]

Lexinfo. URL: http://www.lexinfo.net/ontology/2.0/lexinfo/

[LMF]

Lexical Markup Framework (LMF).
URL: http://www.lexicalmarkupframework.org/

[ONTOLEX]

W3C Ontology Lexica community group.
URL: http://www.w3.org/community/ontolex/

[TR]

Translation Module. URL: http://purl.org/net/translation

[TR_PAPER]

J. Gracia, E. Montiel-Ponsoda, D. Vila-Suero, and G. Aguado-de Cea, *Enabling language resources to expose translations as linked data on the web*, in Proc. of 9th Language Resources and Evaluation Conference (LREC'14), Reykjavik (Iceland), May 2014.

[TRCAT]

OEG Translation Categories. URL: http://purl.org/net/translation-categories

[VOID]

K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao, Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note. March 2011. URL: http://www.w3.org/TR/void/

ReSpec2

4 Converting TBX to RDF

Source: http://www.w3.org/community/bpmlod/wiki/Converting_TBX_to_RDF
Accessed: 14/10/2014

Converting TBX to RDF

From Best Practices for Multilingual Linked Open Data Community Group

Contents

Introduction

This document provides guidlines how to convert terminologies represented in the Term Base eXchange (TBX) into the Resource Description Framework (RDF). TBX is an open standard that has been published by the Localization Industry Standards Association (LISA) (seehere). The standard is identical to ISO standard 30042. This document on the one hand describes the vocabularies that are recommended to be used in doing this conversion and describes the structure of the resulting RDF. It builds on standard W3C vocabularies and other vocabularies that are currently in the process of standardization. The conversion has been implemented in the form of a software package that can be used by anyone (see here).

Selection of vocabularies

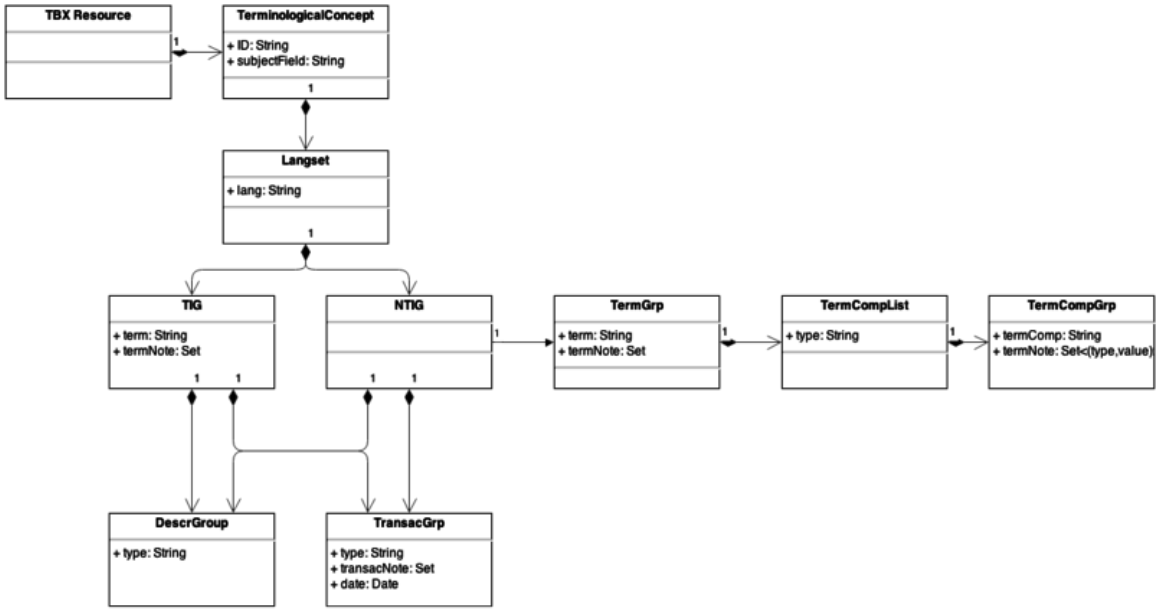
In the following we list the reference models used during the conversion and provide i) in parenthesis the prefix adopted throughout this document; ii) the URL to the model specification.

Model	Prefix	Model reference URL
lemon-ontolex	lemon-ontolex	http://www.w3.org/ns/lemon/ontolex#
SKOS	skos	http://www.w3.org/2004/02/skos/core#
RDF-schema	rdfs	http://www.w3.org/2000/01/rdf-schema#
DCAT	dcat	http://purl.org/dc/terms/
VOID	void	http://rdfs.org/ns/void#
PROV-O: The Prov Ontology	prov	http://www.w3.org/ns/prov#
LIDER TBX Ontology	tbx	http://lider-project.eu/tbx#

We have chosen the lemon-ontolex vocabulary as the backbone of the conversion of TBX into RDF format. lemon-ontolex is a model proposed for representing lexical information relative to ontologies and for linking lexicons and machine-readable dictionaries to the Semantic Web and the Linked Data cloud. The lemon-ontolex vocabulary is currently under discussion by theOntology-Lexicon Community Group that is currently in the process of defining the model. Nevertheless, the model is currently stable enough to build on it.

Technical Description of the Conversion
The TBX Data Model

The following figure summarizes the TBX Data Model as an UML diagram:



- **TBX Resource**: A TBX resource essentially represents a collection of terminological concepts (Terminological Concept), which are represented as XML elements of type termEntry and have a unique ID. In the above XML snippet, there is one terminological concept with ID 2151845. Each terminological concept is described by a set of properties, such as a subject field they belong to.

- **Terminological Concept** (term entry): represents a language-independent concept. Each terminological concept is associated to a LangSet (see below), which can be seen as a set of language-specific Terms that express the Terminological Concept in question.

- **Langset:** A langset is a language-specific containes for all the terms that lexicalize a Terminological Concept in a given language. The Langset contains simple terms, for which no

decompositions is provided (TIG), as well as complex terms for which the decomposition information is provided (NTIG).

- **Term Information Group (TIG):** represents a language-specific term for which no decomposition information is provided.
- **Nesting Term Information Group (NTIG):** represents a language-specific term for which decomposition information is provided.
- **TermGrp:** contains information about a language-specific term including ist morphosyntactic properties; there is one TermGrp for each TIG and NTIG
- **TermCompList:** represents the decomposition of a term
- **TermCompGrp:** represents one component of a term and its morphosyntactic properties
- **DescrGrp:** describes properties of a particular term, in particular different surface forms or describes contexts that document the usage of the term
- **TransGrp/Transaction:** contains information about a transaction that lead to the creation or modification of a term

For a full specification of the TBX data model, please refer to the TBX DTD.

The lemon-ontolex model

To BE DONE (John)

Mapping the TBX Data Model to the ontolex-lemon model

The main data elements described above have been mapped into RDF using the above mentioned vocabularies as follows:

- **TBX Resource:** is not explicitly represented, the whole dataset represents the TBX resources. A TBX resources is thus represented as a void:Dataset. Provenance information is attached, specifying that the data has been converted by the LIDER converter.
- **Terminological Concept:** is represented as a skos:Concept

- **Langset:** A langset is not represented as such in the data. Instead, one ontolex:Lexicon is created for each language for which a Langset is defined. The collection of all the terms for a given language will belong to the corresponding language-specific ontolex:Lexicon

- **TIG/NTIG:** are represented as ontolex:LexicalEntry, no distinction is made between terms with decomposition and terms withouth decomposition; if no decompositions information is available, this is simply omited. In that sense the representation is montonic as the decomposition information can be added later

- **TermGrp:** the information about the morphosyntactic properties of a term is attached to the corresponding ontolex:LexicalEntry. The string enclosed in <term> </term> is assumed to be the ontolex:canonicalForm of the lexical entry in question.

- **TermCompList:** the decomposition of a term is represented using the ontolex:decomp vocabulary, creating a decomp:Component and ontolex:LexicalEntry for each component.

- **TermCompGrp:** the morphosyntactic properties of a component are attached to the corresponding lexical entry that is identified (through decomp:identifies) with the component in question)

- **DescrGrp:** descriptions of the term or context are mapped to appropriate properties of the lexical entry ort he context

- **TransGrp/Transaction:** a transaction that creates or modifies the term is mapped to a tbx:Transaction (a subclass of prov:Activity). Provenance metadata is attached to this entity. The prov:Activity related to the responsible person or agent through prov:wasAssociatedWith; the relation to the responsible Agent is encoded via prov:wasGeneratedBy.

Conversion examples

In this section we provide some examples of excerpts of real TBX documents and how they are converted into RDF following the above guidelines. The example are real example taken from the IATE terminology.

Converting terminological concepts and terms

```
<martif type="TBX-Default" xml:lang="en">
   <martifHeader>
     <fileDesc>
       <sourceDesc>
         <p>This is an excerpt of a TBX file downloaded from the IATE website. Address
any enquiries to iate@cdt.europa.eu.</p>
       </sourceDesc>
```

```
    </fileDesc>
    <encodingDesc>
      <p type="XCSURI">TBXXCS.xcs</p>
    </encodingDesc>
  </martifHeader>
 <text>
    <body>
      <termEntry id="IATE-84">
        <descripGrp>
          <descrip type="subjectField">1011</descrip>
        </descripGrp>
         <langSet xml:lang="de">
          <tig>
            <term>Zuständigkeit der Mitgliedstaaten</term>
            <termNote type="termType">fullForm</termNote>
            <descrip type="reliabilityCode">3</descrip>
          </tig>
         </langSet>
        <langSet xml:lang="en">
          <tig>
            <term>competence of the Member States</term>
            <termNote type="termType">fullForm</termNote>
            <descrip type="reliabilityCode">3</descrip>
          </tig>
        </langSet>
          <langSet xml:lang="es">
          <tig>
            <term>competencias de los Estados miembros</term>
            <termNote type="termType">fullForm</termNote>
            <descrip type="reliabilityCode">3</descrip>
          </tig>
        </langSet>
  </termEntry>
</body>
</text>

         </martif>
```

The resulting RDF would looks as follows:

```
@prefix cc:    <http://creativecommons.org/ns#> . @prefix :
<file:samples/simple1.rdf> . @prefix void:  <http://rdfs.org/ns/void#> . @prefix skos:
<http://www.w3.org/2004/02/skos/core#> . @prefix rdfs:  <http://www.w3.org/2000/01/rdf-
schema#> . @prefix tbx:   <http://tbx2rdf.lider-project.eu/tbx#> . @prefix gr:
<http://purl.org/goodrelations/> . @prefix dct:   <http://purl.org/dc/terms/> . @prefix
rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> . @prefix ontolex:
<http://www.w3.org/ns/ontolex#> . @prefix ldr:   <http://purl.oclc.org/NET/ldr/ns#> .
@prefix odrl:  <http://www.w3.org/ns/odrl/2/> . @prefix dcat:
<http://www.w3.org/ns/dcat#> . @prefix prov:  <http://www.w3.org/ns/prov#> .  :       a
tbx:MartifHeader , dcat:Dataset ;         <http://purl.org/dc/elements/1.1/source>
"This is an excerpt of a TBX file downloaded from the IATE website. Address any enquiries
to iate@cdt.europa.eu." ;         dct:type          "TBX-Default" ;
tbx:encodingDesc  "<p type=\"XCSURI\">TBXXCS.xcs</p>"^^<http://www.w3.org/1999/02/22-rdf-
syntax-ns##XMLLiteral> ;         tbx:sourceDesc    "<sourceDesc><p>This is an excerpt of
```

```
a TBX file downloaded from the IATE website. Address any enquiries to
iate@cdt.europa.eu.</p></sourceDesc>"^^<http://www.w3.org/1999/02/22-rdf-syntax-
ns##XMLLiteral> .  :Lexicon_de  a  ontolex:Lexicon ;  ontolex:entry
 :LexicalEntry-2d10ceeb-2db2-4b9d-9ffb-bfc0b28a63ab ;  ontolex:language  "de" .
:Lexicon_es  a  ontolex:Lexicon ;  ontolex:entry  :LexicalEntry-
b91ad96c-a45f-403a-969c-2a88a49eef9f ;  ontolex:language  "es" .  :Lexicon_en  a
ontolex:Lexicon ;  ontolex:entry  :LexicalEntry-6759a1c9-2b9a-4c5f-9a95-
9d744ea0f3da ;  ontolex:language  "en" .  :LexicalEntry-6759a1c9-2b9a-4c5f-9a95-
9d744ea0f3da  a  ontolex:LexicalEntry ;
tbx:reliabilityCode  "3"^^tbx:reliabilityCode ;  tbx:termType
tbx:fullForm ;  ontolex:canonicalForm  :LexicalEntry-6759a1c9-2b9a-4c5f-9a95-
9d744ea0f3da-CanonicalForm ;  ontolex:language  "en" ;  ontolex:sense
 :LexicalEntry-6759a1c9-2b9a-4c5f-9a95-9d744ea0f3da-Sense .  :LexicalEntry-b91ad96c-a45f-
403a-969c-2a88a49eef9f-CanonicalForm  ontolex:writtenRep  "competencias de los
Estados miembros"@es .  :LexicalEntry-2d10ceeb-2db2-4b9d-9ffb-bfc0b28a63ab-Sense
ontolex:reference  :Term-96ecadd1-e352-4e87-927a-6f3ab5a550ba .  :Term-96ecadd1-e352-
4e87-927a-6f3ab5a550ba  a
<<http://www.w3.org/2004/02/skos/core#Concept> ;  tbx:subjectField
"1011"^^tbx:subjectField .  :LexicalEntry-b91ad96c-a45f-403a-969c-2a88a49eef9f-Sense
ontolex:reference  :Term-96ecadd1-e352-4e87-927a-6f3ab5a550ba .  :LexicalEntry-2d10ceeb-
2db2-4b9d-9ffb-bfc0b28a63ab  a  ontolex:LexicalEntry ;
tbx:reliabilityCode  "3"^^tbx:reliabilityCode ;  tbx:termType
tbx:fullForm ;  ontolex:canonicalForm  :LexicalEntry-2d10ceeb-2db2-4b9d-9ffb-
bfc0b28a63ab-CanonicalForm ;  ontolex:language  "de" ;  ontolex:sense
 :LexicalEntry-2d10ceeb-2db2-4b9d-9ffb-bfc0b28a63ab-Sense .  :LexicalEntry-6759a1c9-2b9a-
4c5f-9a95-9d744ea0f3da-Sense  ontolex:reference  :Term-96ecadd1-e352-4e87-927a-
6f3ab5a550ba .  :LexicalEntry-2d10ceeb-2db2-4b9d-9ffb-bfc0b28a63ab-CanonicalForm
ontolex:writtenRep  "Zuständigkeit der Mitgliedstaaten"@de .  :LexicalEntry-b91ad96c-
a45f-403a-969c-2a88a49eef9f  a  ontolex:LexicalEntry ;
tbx:reliabilityCode  "3"^^tbx:reliabilityCode ;  tbx:termType
tbx:fullForm ;  ontolex:canonicalForm  :LexicalEntry-b91ad96c-a45f-403a-969c-
2a88a49eef9f-CanonicalForm ;  ontolex:language  "es" ;  ontolex:sense
 :LexicalEntry-b91ad96c-a45f-403a-969c-2a88a49eef9f-Sense .  :LexicalEntry-6759a1c9-
2b9a-4c5f-9a95-9d744ea0f3da-CanonicalForm  ontolex:writtenRep  "competence of the
Member States"@en .
```

Note that the terminology entry has been represented as a skos:Concept, which gets assigned a subjecField (1011 in this case). Further, there is one lexicon object for each of the three languages (en,de,es) containing one lexical entry for the corresponding term. A sense has been introduced to represent the meaning of each of these lexical entry a referring to the correspoding terminology concept.

Converting term decomposition information

The following example extends the previous one by adding decomposition information for one term:

```
<?xml version="1.0" encoding="utf-8"?> <martif type="TBX-Default" xml:lang="en">
<martifHeader>     <fileDesc>       <sourceDesc>         <p>This is a TBX file downloaded
from the IATE website. Address any enquiries to iate@cdt.europa.eu.</p>
</sourceDesc>     </fileDesc>     <encodingDesc>       <p type="XCSURI">TBXXCS.xcs</p>
</encodingDesc>   </martifHeader>   <text>   <body><termEntry id="IATE-14">   <descrip
type="subjectField">1011</descrip>    <langSet xml:lang="en">     <ntig>       <termGrp>       <term>compete
<termComp>competence</termComp>             <termNote type="partOfSpeech">noun</termNote>
<termNote type="grammaticalNumber">singular</termNote>           </termCompGrp>
<termCompGrp>             <termComp>of</termComp>             <termNote
type="partOfSpeech">other</termNote>           </termCompGrp>           <termCompGrp>
<termComp>the</termComp>             <termNote type="partOfSpeech">other</termNote>
</termCompGrp>        <termCompGrp>             <termComp>Member</termComp>
<termNote type="partOfSpeech">noun</termNote>        <termNote
type="grammaticalNumber">singular</termNote>               </termCompGrp>
<termCompGrp>             <termComp>States</termComp>             <termNote
type="partOfSpeech">noun</termNote>             <termNote
type="grammaticalNumber">plural</termNote>           </termCompGrp>
</termCompList>       </termGrp>     </ntig>    </langSet> </termEntry> </body> </text>
</martif>
```

The resulting RDF would look as follows:

```
@prefix cc:    <http://creativecommons.org/ns#> . @prefix :
<file:samples/simple_with_decomposition.rdf> . @prefix void:  <http://rdfs.org/ns/void#>
. @prefix skos:  <http://www.w3.org/2004/02/skos/core#> . @prefix rdfs:
<http://www.w3.org/2000/01/rdf-schema#> . @prefix tbx:   <http://tbx2rdf.lider-
project.eu/tbx#> . @prefix gr:    <http://purl.org/goodrelations/> . @prefix dct:
<http://purl.org/dc/terms/> . @prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-
ns#> . @prefix ontolex: <http://www.w3.org/ns/ontolex#> . @prefix ldr:
<http://purl.oclc.org/NET/ldr/ns#> . @prefix odrl:  <http://www.w3.org/ns/odrl/2/> .
@prefix dcat:  <http://www.w3.org/ns/dcat#> . @prefix prov:  <http://www.w3.org/ns/prov#>
.  :       a                 tbx:MartifHeader ;
<http://purl.org/dc/elements/1.1/source>                 "This is a TBX file downloaded
from the IATE website. Address any enquiries to iate@cdt.europa.eu." ;         dct:type
"TBX-Default" ;         tbx:encodingDesc  "<p
```

```
type=\"XCSURI\">TBXXCS.xcs</p>"^^<http://www.w3.org/1999/02/22-rdf-syntax-
ns##XMLLiteral> ;         tbx:sourceDesc    "<sourceDesc><p>This is a TBX file downloaded
from the IATE website. Address any enquiries to
iate@cdt.europa.eu.</p></sourceDesc>"^^<http://www.w3.org/1999/02/22-rdf-syntax-
ns##XMLLiteral> . :Lexicon_en  a            ontolex:Lexicon ;         ontolex:entry
 :LexicalEntry-2228c6fe-a23f-4ff7-af11-072cec346d8a ;         ontolex:language  "en" .
:Term-6104f578-8359-4004-a904-f186a5c01fc1         a
<<http://www.w3.org/2004/02/skos/core#Concept> ;         tbx:subjectField  "1011"@en .
:LexicalEntry-2228c6fe-a23f-4ff7-af11-072cec346d8a-Sense         ontolex:reference
 :Term-6104f578-8359-4004-a904-f186a5c01fc1 .  :TermComp-1aca99b1-ab82-4924-9dea-
ef5c581408bd         a                 ontolex:LexicalEntry ;         rdfs:label
"the"@en ;         tbx:partOfSpeech  tbx:other ;
<http://www.w3.org/ns/decomp#identifies>                 :TermComp-1aca99b1-ab82-4924-
9dea-ef5c581408bd .  :LexicalEntry-2228c6fe-a23f-4ff7-af11-072cec346d8a-CanonicalForm
ontolex:writtenRep  "competence of the Member States"@en .  :TermComp-0ab24893-3f57-4f0c-
8f53-42dc12b79107         a                       ontolex:LexicalEntry ;
rdfs:label             "Member"@en ;         tbx:grammaticalNumber  tbx:singular ;
tbx:partOfSpeech       tbx:noun ;         <http://www.w3.org/ns/decomp#identifies>
 :TermComp-0ab24893-3f57-4f0c-8f53-42dc12b79107 .  :TermComp-e9f4b191-9356-49bb-bde1-
e3d61b2f2da5         a                       ontolex:LexicalEntry ;         rdfs:label
"competence"@en ;         tbx:grammaticalNumber  tbx:singular ;         tbx:partOfSpeech
tbx:noun ;         <http://www.w3.org/ns/decomp#identifies>                 :TermComp-
e9f4b191-9356-49bb-bde1-e3d61b2f2da5 .  :TermCompList-581d978b-ab24-44a9-95bc-
9b5f5e21e21b         <http://www.w3.org/ns/decomp#constituent>                 :TermComp-
0bff3192-f6db-4b62-805b-63ba0c23bd72 , :TermComp-0ab24893-3f57-4f0c-8f53-42dc12b79107
, :TermComp-1aca99b1-ab82-4924-9dea-ef5c581408bd , :TermComp-d7d1d67c-c4d3-4465-94ab-
0b5866f55735 , :TermComp-e9f4b191-9356-49bb-bde1-e3d61b2f2da5 ;
<http://www.w3.org/ns/decomp#identifies>                 :LexicalEntry-2228c6fe-a23f-
4ff7-af11-072cec346d8a .  :LexicalEntry-2228c6fe-a23f-4ff7-af11-072cec346d8a         a
ontolex:LexicalEntry ;         tbx:termType           tbx:fullForm ;
ontolex:canonicalForm  :LexicalEntry-2228c6fe-a23f-4ff7-af11-072cec346d8a-CanonicalForm ;
ontolex:language       "en" ;         ontolex:sense          :LexicalEntry-2228c6fe-a23f-
4ff7-af11-072cec346d8a-Sense .  :TermComp-0bff3192-f6db-4b62-805b-63ba0c23bd72         a
ontolex:LexicalEntry ;         rdfs:label             "States"@en ;
tbx:grammaticalNumber  tbx:plural ;         tbx:partOfSpeech       tbx:noun ;
<http://www.w3.org/ns/decomp#identifies>                 :TermComp-0bff3192-f6db-4b62-
805b-63ba0c23bd72 .  :TermComp-d7d1d67c-c4d3-4465-94ab-0b5866f55735         a
ontolex:LexicalEntry ;         rdfs:label        "of"@en ;         tbx:partOfSpeech
tbx:other ;         <http://www.w3.org/ns/decomp#identifies>                 :TermComp-
d7d1d67c-c4d3-4465-94ab-0b5866f55735 .
```

The above example shows how one terminological entry with one term in English has been transformed into ontolex-based RDF. Note that in this case only one lexicon has been created. Further, to repersent the composition information, a TermCompList object has been created that lists all constituents of the complex terms corresponding to each of the words that make up the compound. Each constituent is associated to a corresponding lexical entry that links to the corresponding constituent through the identifies relation.

Converting transaction information

The following example adds transaction information to one term:

```
<?xml version="1.0" encoding="utf-8"?> <martif type="TBX-Default" xml:lang="en">
<martifHeader>     <fileDesc>       <sourceDesc>         <p>This is an excerpt of a TBX
file downloaded from the IATE website. Address any enquiries to iate@cdt.europa.eu.</p>
</sourceDesc>     </fileDesc>     <encodingDesc>       <p type="XCSURI">TBXXCS.xcs</p>
</encodingDesc>   </martifHeader>   <text>   <body><termEntry id="IATE-14">   <descrip
type="subjectField">1011</descrip>    <langSet xml:lang="en">     <ntig>       <termGrp>
 <term>competence of the Member States</term>          <termNote
type="termType">fullForm</termNote>          <descrip type="reliabilityCode">3</descrip>
<termCompList type="lemma">           <termCompGrp>
<termComp>competence</termComp>             <termNote type="partOfSpeech">noun</termNote>
<termNote type="grammaticalNumber">singular</termNote>           </termCompGrp>
<termCompGrp>             <termComp>of</termComp>             <termNote
type="partOfSpeech">other</termNote>           </termCompGrp>           <termCompGrp>
<termComp>the</termComp>             <termNote type="partOfSpeech">other</termNote>
</termCompGrp>        <termCompGrp>             <termComp>Member</termComp>
<termNote type="partOfSpeech">noun</termNote>        <termNote
type="grammaticalNumber">singular</termNote>               </termCompGrp>
<termCompGrp>             <termComp>States</termComp>             <termNote
type="partOfSpeech">noun</termNote>             <termNote
type="grammaticalNumber">plural</termNote>           </termCompGrp>
</termCompList>       </termGrp>       <transacGrp>         <transac
type="transactionType">origination</transac>         <transacNote
type="responsibility">PC</transacNote>         <date>2014-05-08</date>
</transacGrp>       <transacGrp>         <transac
type="transactionType">approval</transac>         <transacNote
type="responsibility">PC</transacNote>         <date>2014-05-16T14:50:42.018Z</date>
</transacGrp>       <admin type="status">approved</admin>       <transacGrp>
<transac type="transactionType">modification</transac>         <transacNote
type="responsibility">PC</transacNote>         <date>2014-05-16T14:59:00.814Z</date>
</transacGrp>     </ntig>    </langSet> </termEntry> </body> </text> </martif>
```

The resulting RDF would look as follows:

```
@prefix cc:    <http://creativecommons.org/ns#> . @prefix :
<file:samples/simple_with_decomp_trans.rdf> . @prefix void:  <http://rdfs.org/ns/void#> .
@prefix skos:  <http://www.w3.org/2004/02/skos/core#> . @prefix rdfs:
<http://www.w3.org/2000/01/rdf-schema#> . @prefix tbx:   <http://tbx2rdf.lider-
project.eu/tbx#> . @prefix gr:    <http://purl.org/goodrelations/> . @prefix dct:
<http://purl.org/dc/terms/> . @prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-
ns#> . @prefix ontolex: <http://www.w3.org/ns/ontolex#> . @prefix ldr:
<http://purl.oclc.org/NET/ldr/ns#> . @prefix odrl:  <http://www.w3.org/ns/odrl/2/> .
@prefix dcat:  <http://www.w3.org/ns/dcat#> . @prefix prov:  <http://www.w3.org/ns/prov#>
.  :       a                tbx:MartifHeader , dcat:Dataset ;
<http://purl.org/dc/elements/1.1/source>                "This is an excerpt of a TBX
file downloaded from the IATE website. Address any enquiries to iate@cdt.europa.eu." ;
dct:type          "TBX-Default" ;         tbx:encodingDesc  "<p
type=\"XCSURI\">TBXXCS.xcs</p>"^^<http://www.w3.org/1999/02/22-rdf-syntax-
ns##XMLLiteral> ;         tbx:sourceDesc    "<sourceDesc><p>This is an excerpt of a TBX
file downloaded from the IATE website. Address any enquiries to
iate@cdt.europa.eu.</p></sourceDesc>"^^<http://www.w3.org/1999/02/22-rdf-syntax-
ns##XMLLiteral> .  :Lexicon_en  a            ontolex:Lexicon ;         ontolex:entry
 :LexicalEntry-7fd6e53c-1da9-4419-a5e1-3e52d8c8b4b6 ;         ontolex:language  "en" .
:LexicalEntry-7fd6e53c-1da9-4419-a5e1-3e52d8c8b4b6         a
ontolex:LexicalEntry ;         <http://myproperty/status>
"approved"^^<http://myproperty/status> ;         tbx:termType
tbx:fullForm ;         tbx:transaction             :TransacGrp-306a67e3-cf0d-4c20-af4d-
38e1f42d1487 , :TransacGrp-7f2e3d2e-7639-4ff1-afe2-e131078b1334 , :TransacGrp-87989e8f-
3e53-4c54-974f-72bdee3e25fe ;         ontolex:canonicalForm       :LexicalEntry-7fd6e53c-
1da9-4419-a5e1-3e52d8c8b4b6-CanonicalForm ;         ontolex:language            "en" ;
ontolex:sense               :LexicalEntry-7fd6e53c-1da9-4419-a5e1-3e52d8c8b4b6-Sense .
:LexicalEntry-7fd6e53c-1da9-4419-a5e1-3e52d8c8b4b6-CanonicalForm
ontolex:writtenRep  "competence of the Member States"@en .  :TermCompList-319356e5-ae82-
4d2d-9ef1-35bfefee9c57         <http://www.w3.org/ns/decomp#constituent>
 :TermComp-5de261a7-2092-4b22-8a08-4b450fb9a1fe , :TermComp-0b0ad58e-e2ba-42eb-9c99-
415fe6b6fc18 , :TermComp-c94a5d19-3aa8-4ff0-8f6a-38b4b6c15d59 , :TermComp-4ba2b9da-cda9-
4fd1-abb6-eec7344ee763 , :TermComp-c0b624f0-5cc2-4f14-9f63-7e6270ed52d4 ;
<http://www.w3.org/ns/decomp#identifies>                :LexicalEntry-7fd6e53c-1da9-
4419-a5e1-3e52d8c8b4b6 .  :TermComp-0b0ad58e-e2ba-42eb-9c99-415fe6b6fc18         a
ontolex:LexicalEntry ;         rdfs:label             "Member"@en ;
tbx:grammaticalNumber  tbx:singular ;         tbx:partOfSpeech       tbx:noun ;
<http://www.w3.org/ns/decomp#identifies>                :TermComp-0b0ad58e-e2ba-42eb-
```

```
9c99-415fe6b6fc18 .  :TransacGrp-7f2e3d2e-7639-4ff1-afe2-e131078b1334         a
prov:Activity , tbx:transaction ;         tbx:transactionType
"approval"^^tbx:transactionType ;         prov:endedAtTime        "2014-05-
16T14:50:42.018Z"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
prov:wasAssociatedWith  :TransacNote-41238152-7acf-4353-8479-46e3ff068757 .
:TransacNote-41238152-7acf-4353-8479-46e3ff068757         a           prov:Agent ;
rdfs:label  "PC" .  :TermComp-c94a5d19-3aa8-4ff0-8f6a-38b4b6c15d59         a
ontolex:LexicalEntry ;         rdfs:label        "the"@en ;         tbx:partOfSpeech
tbx:other ;         <http://www.w3.org/ns/decomp#identifies>                 :TermComp-
c94a5d19-3aa8-4ff0-8f6a-38b4b6c15d59 .  :TransacGrp-306a67e3-cf0d-4c20-af4d-38e1f42d1487
a                       prov:Activity , tbx:transaction ;         tbx:transactionType
"modification"^^tbx:transactionType ;         prov:endedAtTime        "2014-05-
16T14:59:00.814Z"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
prov:wasAssociatedWith  :TransacNote-2e0fd885-5624-4e6f-b52d-6e1ebcef9ada .
:TransacNote-2e0fd885-5624-4e6f-b52d-6e1ebcef9ada         a           prov:Agent ;
rdfs:label  "PC" .  :TermComp-5de261a7-2092-4b22-8a08-4b450fb9a1fe         a
ontolex:LexicalEntry ;         rdfs:label             "States"@en ;
tbx:grammaticalNumber  tbx:plural ;         tbx:partOfSpeech       tbx:noun ;
<http://www.w3.org/ns/decomp#identifies>                 :TermComp-5de261a7-2092-4b22-
8a08-4b450fb9a1fe .  :LexicalEntry-7fd6e53c-1da9-4419-a5e1-3e52d8c8b4b6-Sense
ontolex:reference  :Term-77d008db-5141-434c-a25c-7b34aaac76d4 .  :TermComp-4ba2b9da-cda9-
4fd1-abb6-eec7344ee763         a                 ontolex:LexicalEntry ;
rdfs:label        "of"@en ;         tbx:partOfSpeech  tbx:other ;
<http://www.w3.org/ns/decomp#identifies>                 :TermComp-4ba2b9da-cda9-4fd1-
abb6-eec7344ee763 .  :TransacGrp-87989e8f-3e53-4c54-974f-72bdee3e25fe         a
prov:Activity , tbx:transaction ;         tbx:transactionType
"origination"^^tbx:transactionType ;         prov:endedAtTime        "2014-05-
08"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;         prov:wasAssociatedWith
 :TransacNote-4f64b4fe-0478-4a9e-bb7a-20a54ce17242 .  :TransacNote-4f64b4fe-0478-4a9e-
bb7a-20a54ce17242         a           prov:Agent ;         rdfs:label  "PC" .  :Term-
77d008db-5141-434c-a25c-7b34aaac76d4         a
<<http://www.w3.org/2004/02/skos/core#Concept> ;         tbx:subjectField
"1011"^^tbx:subjectField .  :TermComp-c0b624f0-5cc2-4f14-9f63-7e6270ed52d4         a
ontolex:LexicalEntry ;         rdfs:label             "competence"@en ;
tbx:grammaticalNumber  tbx:singular ;         tbx:partOfSpeech       tbx:noun ;
<http://www.w3.org/ns/decomp#identifies>                 :TermComp-c0b624f0-5cc2-4f14-
9f63-7e6270ed52d4 .
```

This example extends the previous example by transactions. Note that each transaction element has been mapped to a **tbx:Transaction** element and declared as having type **prov:Activity**. The transction type as originating form the data is faitfhfully represented. The data of the transaction is

mapped to the property prov:endedAtTime and the responsible field has been mapped to the property prov:wasAssociatedWith. The date has been represented as the property **prov:endedAtDate** and the associated responsible has been represented via the property **prov:isAssociatedWith**. The admin status "approved" has been mapped to a proprietary property <http://myproperty/status> as it is not defined in the TBX standard.

Proof-of-Concept

As a proof-of-concept for the conversion, we have converted the IATE (InterActive Terminology of Europe) into RDF format. The data is available for download here. The download file contains about 8 million terms in 24 official EU languages.

Data querying

In order to show how the model can be used, we give a number of queries that demonstrate data access to the model.

All of these queries are incorrect and need to be fixed!

Indicate the number of terminological concepts in the resource

```
select distinct count(?Concept) from <http://tbx2rdf.lider-project.eu/> where {
 ?Concept a <http://www.w3.org/2004/02/skos/core#Concept>  }
```

Try it out

Indicates all the languages of terms included in the resource

```
select distinct ?language from <http://tbx2rdf.lider-project.eu/> where {    ?c
<http://www.w3.org/ns/lemon/ontolex#language> ?language  }
```

Try it out

Retrieve all the canonical forms for a given terminological concept in all languages, with indication of the language tag

```
PREFIX ontolex: <http://www.w3.org/ns/ontolex#>   select distinct ?cf from
<http://tbx2rdf.lider-project.eu/> where {    ?sense ontolex:reference
<http://tbx2rdf.lider-project.eu#Term-0327f3e6-6b69-4ba1-81f8-6425b858a0af> .   ?entry
ontolex:sense ?sense ;     ontolex:canonicalForm ?form .   ?form ontolex:writtenRep ?cf .
} limit 100
```

Try it out

Retrieve all terms by subject field

```
PREFIX ontolex: <http://www.w3.org/ns/ontolex#>   select distinct ?c ?cf from
<http://tbx2rdf.lider-project.eu/> where {    ?c a
<http://www.w3.org/2004/02/skos/core#Concept> ;      <http://tbx2rdf.lider-
project.eu/tbx#subjectField> "1011"^^<http://tbx2rdf.lider-project.eu/tbx#subjectField> .
 ?sense ontolex:reference ?c .   ?entry ontolex:sense ?sense ;
ontolex:canonicalForm ?form .   ?form ontolex:writtenRep ?cf . } LIMIT 100
```

Try it out

Implementation

A converter has been implemented to map TBX/XML input into RDF using the vocabularies described above. The converter has been implemented as a Java program that reads in the document and builds the DOM tree. The DOM tree is traversed and elements are mapped to appropriate object-oriented datastructures. These datastructures are then serialized as RDF. The code is available as GitHub project tbx2rdf. Further, a web service that implements the conversion functionality is available here: http://tbx2rdf.appspot.com/. As additional input to the program, a file can be provided that contains mappings of specific XML elements and attributes used in the TBX document to URIs representing properties. If no file is specified the default file „default.mappings“ is used. This option is only available when directly executing the Java program, not via the Web service.

A service for converting TBX to RDF is available here: http://tbx2rdf.lider-project.eu/converter

References

TBX Standard as published by LISA

tbx2rdf GitHub Project

5 NIF Web Services

Source: https://www.w3.org/community/bpmlod/wiki/NIF_Web_Services
Accessed: 14/10/2014

NIF Web Services

From Best Practices for Multilingual Linked Open Data Community Group

Contents

Introduction

This document describes best practices to follow for the implementation of RESTful NLP web services that rely on the NLP Interchange Format (NIF). „NIF is an RDF/OWL-based format that aims to achieve interoperability between NLP tools language resources and annotations.“ As a proof-of-concept, we have implemented NIF wrappes for the Stanford POS tagger and Stanford parser.

Natural Language Processing Interchange Format (NIF)

NIF is an RDF-based format. The classes to represent linguistic data are defined in the NIF Core Ontology. All ontology classes are derived from the main class nif:String which respresents strings of Unicode characters. One important subclass of nif:String is nif:Context. It represents a text in its entirety and holds the characters of this text in the nif:isString property. There are several classes (e.g. nif:Word, nif:Phrase, nif:Sentence) for representing partitions of a text, their choice depends on the unit of annotation. All such subunits have a property nif:referenceContext pointing to their respective

nif:Context instance. Furthermore, their position inside the context is specified using the nif:beginIndex and nif:endIndex properties. The actual substring represented by these units can be specified using the nif:anchorOf property. Annotations like POS tags or relation types (see below) can be added as properties to the respective nif.String objects. NIF individuals are identified by URIs following a nif:URIScheme which restricts the URI's syntax. E.g. a URI following RFC 5147 consists of a prefix string followed by „#char=x,y“, where x and y are the start and end positions of the string in its context. For nif:Context URIs y can be omitted or set to the total number of characters in the text.

Recommended service parameters

NIF services should conform to the NIF 2.0 public API specification. The following parameters are supported by a specification compliant service. *Required* parameters need to be specified by the user in order for the service to function. *Optional* parameters can be omitted, in which case default values are used by the service.

Required:

- input (i): The input to be processed by the service.

Optional:

- informat (f): The format in which the input is given. Supported argument values are text, turtle (default) and json-ld.
- intype (t): Specifies how the input is retrieved. Supported argument values are direct (default), file and url.
- outformat (o): The format in which the output will be serialized. Supported argument values are turtle (default) and json-ld.
- urischeme (u): the URI scheme the service must use to create new URIs
- prefix (p): the service must use this as the prefix part of new URIs. A UUID will be generated if no prefix is specified

Furthermore, we recommend to implement a parameter *info* which, according to the NIF API specification can be used to output all implemented parameters if info=true. In addition to that, we recommend to output supported parameters and default values as well.

Further recommended parameters which not part of the NIF API specification are the following:

- verbosity (v): Accepting two values: true and false. True returns full output in NIF format, while false returns only the triples added to the data
- model (m): the path/url of a trained model to be used by the service, a default model should be used if no model is specified
- language (l): a parameter specifying the language of the input, default is English

Log messages

NIF services should generate log messages in RDF format using the RDF Logging Ontology. An rlog message is of type rlog:entry and should contain the properties rlog:level, rlog:date and rlog:message. We recommend to generate a log entry in the following cases:

- If no input is specified. Log level should be rlog:FATAL.
- If the input is given as file or url but couldn't be retrieved by the service. Log level should be rlog:FATAL.
- If a parameter value isn't supported by the service. Log level should be rlog:FATAL.
- If an optional parameter is omitted. Log level should be rlog:WARN. The message should state the default value being used.

Example Implementations
Wrapping the Stanford POS Tagger

Our web service wrapping the Stanford POS tagger can be invoked via curl using the following example call.

Example call 1:

```
curl vtentacle.techfak.uni-bielefeld.de/~bsiemone/index.php/NifStanfordPOSTagger -d
f="text" -d i="This is a sample sentence"
```

If the input is given as plain text like in the above example, an RDF model is constructed containing a nif:Context element with the input text in its nif:isString property and one nif:Word element for each word in the input. If the input is already in NIF format, it is expected to contain at least one nif:Context element. Right now all input elements except the instances of nif:Context are ignored. It would be possible to tag instances of nif:Sentence or sets of nif:Word though. The service then reads the nif:isString values of all nif:Context elements found in the input and passes them to the Stanford NLP tools where they are tokenized and POS tagged. Each word is annotated by adding a nif:posTag property with the POS tag as a literal value to the corresponding nif:Word element. If a nif:Word instance is missing in the input it will be created.

The example output of the service can be found here:

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .  <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=0,4>         a                      nif:RFC5147String , nif:Word ;
nif:anchorOf          "This"^^xsd:string ;         nif:beginIndex        "0"^^xsd:int ;
```

```
nif:endIndex "4"^^xsd:int ; nif:posTag "DT"^^xsd:string ;
nif:referenceContext <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=0,25> . <a3ecec2a-66dc-
45af-aa0f-0723138e3de1#char=5,7> a nif:RFC5147String ,
nif:Word ; nif:anchorOf "is"^^xsd:string ; nif:beginIndex
"5"^^xsd:int ; nif:endIndex "7"^^xsd:int ; nif:posTag
"VBZ"^^xsd:string ; nif:referenceContext <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=0,25> . <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=0,25> a
nif:Context , nif:RFC5147String , nif:String ; nif:isString "This is a sample
sentence"^^xsd:string . <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=10,16> a
nif:RFC5147String , nif:Word ; nif:anchorOf "sample"^^xsd:string ;
nif:beginIndex "10"^^xsd:int ; nif:endIndex "16"^^xsd:int ;
nif:posTag "NN"^^xsd:string ; nif:referenceContext <a3ecec2a-66dc-
45af-aa0f-0723138e3de1#char=0,25> . <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=8,9>
a nif:RFC5147String , nif:Word ; nif:anchorOf
"a"^^xsd:string ; nif:beginIndex "8"^^xsd:int ; nif:endIndex
"9"^^xsd:int ; nif:posTag "DT"^^xsd:string ;
nif:referenceContext <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=0,25> . <a3ecec2a-66dc-
45af-aa0f-0723138e3de1#char=17,25> a nif:RFC5147String ,
nif:Word ; nif:anchorOf "sentence"^^xsd:string ; nif:beginIndex
"17"^^xsd:int ; nif:endIndex "25"^^xsd:int ; nif:posTag
"NN"^^xsd:string ; nif:referenceContext <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=0,25> .
```

Wrapping the Stanford Parser

Our web service wrapping the Stanford dependency parser can be invoked via curl using the following example call where the input is assumed to be given in a turtle file called input.tll.

Example call 2:

```
curl vtentacle.techfak.uni-bielefeld.de/~bsiemone/index.php/NifStanfordParser -d
i="input.ttl"
```

The service can be used to parse input that is already POS tagged. I.e. it expects the input to be in NIF format and contain a) at least one nif:Context element b) one nif:Word element for each word in the nif:isString property of its context containing a POS annotation in nif:posTag and the represented string in nif:anchorOf. The words are ordered by context (using nif:referenceContext) and position (using nif:beginIndex) in order to reconstruct the original texts. The service then passes the annotated input to the Stanford parser. For each dependency relation of the parse a nif:dependency property is added to the relation's head with the URI of the dependent word as object. As a word can only have one head, the type of the relation is annotated in the nif:dependencyRelationType property of the dependent word (as a literal).

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .  <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=0,25>         a             nif:String , nif:RFC5147String ,
nif:Context ;         nif:isString  "This is a sample sentence"^^xsd:string .  <a3ecec2a-
66dc-45af-aa0f-0723138e3de1#char=8,9>         a                          nif:Word ,
nif:RFC5147String ;         nif:anchorOf               "a"^^xsd:string ;
nif:beginIndex              "8"^^xsd:int ;         nif:dependencyRelationType
"dep"^^xsd:string ;         nif:endIndex               "9"^^xsd:int ;         nif:posTag
"DT"^^xsd:string ;         nif:referenceContext        <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=0,25> .  <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=10,16>         a
nif:Word , nif:RFC5147String ;         nif:anchorOf               "sample"^^xsd:string ;
nif:beginIndex              "10"^^xsd:int ;         nif:dependencyRelationType
"nn"^^xsd:string ;         nif:endIndex               "16"^^xsd:int ;         nif:posTag
"NN"^^xsd:string ;         nif:referenceContext        <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=0,25> .  <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=0,4>         a
nif:Word , nif:RFC5147String ;         nif:anchorOf               "This"^^xsd:string ;
nif:beginIndex              "0"^^xsd:int ;         nif:dependencyRelationType
"nsubj"^^xsd:string ;         nif:endIndex               "4"^^xsd:int ;
nif:posTag                  "DT"^^xsd:string ;         nif:referenceContext
<a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=0,25> .  <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=17,25>         a                     nif:Word , nif:RFC5147String ;
nif:anchorOf          "sentence"^^xsd:string ;         nif:beginIndex
"17"^^xsd:int ;         nif:dependency        <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=8,9> , <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=5,7> , <a3ecec2a-
66dc-45af-aa0f-0723138e3de1#char=26,27> , <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=0,4>
, <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=10,16> ;         nif:endIndex
"25"^^xsd:int ;         nif:posTag            "NN"^^xsd:string ;
nif:referenceContext  <a3ecec2a-66dc-45af-aa0f-0723138e3de1#char=0,25> .  <a3ecec2a-66dc-
45af-aa0f-0723138e3de1#char=5,7>         a                          nif:Word ,
nif:RFC5147String ;         nif:anchorOf               "is"^^xsd:string ;
nif:beginIndex              "5"^^xsd:int ;         nif:dependencyRelationType
"aux"^^xsd:string ;         nif:endIndex               "7"^^xsd:int ;         nif:posTag
"VBZ"^^xsd:string ;         nif:referenceContext        <a3ecec2a-66dc-45af-aa0f-
0723138e3de1#char=0,25> .
```

Chaining

As one of the services described above (the tagger) produces output the other one (the parser) relies on, they can be used to demonstrate the integration of NIF compliant NLP services. The following nested call combines example calls 1 and 2. It invokes the tagger which produces the output of

example call 1 and passes this POS annotated NIF data to the parser. The output is the same as in example call 2.

Example call 3:

```
curl vtentacle.techfak.uni-bielefeld.de/~bsiemone/index.php/NifStanfordParser  -d
i=$(curl vtentacle.techfak.uni-bielefeld.de/~bsiemone/index.php/NifStanfordPOSTagger -d
f="text" -d i="This is a sample sentence")
```

References

[NIF Core Ontology]

[NIF 2.0 public API specification]

[RLOG - an RDF Logging Ontology]