Guidelines for transforming terminological resources into RDF

In this document we describe guidelines for transforming terminological resources. We focus on TermBase eXchange (TBX) format as this is a widely used model that has been standardized by ISO (ISO 30042). These guidelines should not be regarded as a complete specification for how to convert TBX documents into RDF/ Linked Data, but rather as a set of best practices and guidelines to follow when doing so in order to maximize vocabulary reuse and thus interoperability with other resources:

Vocabularies

The vocabularies that we propose to use in the translation are the following ones; reused classes / properties are indicated:

Name	Abbr.	URL	Reused Elements
Resource Description Framework	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns	**Properties:** type, _1, _2, _3
Resource Description Framework Schema	rdfs	http://www.w3.org/2000/01/rdf-schema	
Dublin Code	dc	http://purl.org/dc/terms	**Properties:** source, creator
SKOS	skos	http://www.w3.org/2004/02/skos/core	**Classes:** Concept
Provenance	prov	http://www.w3.org/ns/prov	**Classes:** Activity, Entity **Properties:** endedAtTime wasAssociatedWith wasGeneratedBy
Lexicon Model for Ontologies – Core Module	ontolex	http://www.w3.org/ns/ontolex	**Classes:** Lexicon, LexicalEntry, LexicalSense **Properties:** language, canonicalForm lexicalizedBy, entry, definition, writtenRep, otherForm, sense
Lexicon Model	decomp	http://www.w3.org/ns/lemon/decomp	**Classes:**

for Ontologies – Decomposition Module			Component **Properties:** constituent identifies
Vocabulary of Interlinked Datasets	void	http://www.w3.org/TR/void/	**Classes:** Dataset

Data Sample

The following XML Code shows a data sample provided by TILDE:

```
<?xml version="1.0" encoding="utf-8"?>
<martif type="TBX" xml:lang="en">
  <martifHeader>
    <fileDesc>
      <sourceDesc>
        <p>LOD_experiment</p>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <p type="XCSURI">http://www.ttt.org/oscarstandards/tbx/TBXXCS.xcs</p>
    </encodingDesc>
  </martifHeader>
  <text>
    <body><termEntry id="2151845">
  <descrip type="subjectField">TaaS-1500</descrip>
  <langSet xml:lang="lv" xmlns:xml="http://www.w3.org/XML/1998/namespace">
    <ntig>
      <termGrp>
        <term>globālais tīmeklis</term>
        <termCompList type="lemma" />
      </termGrp>
      <xref target="http://www.eurotermbank.com/Collection.aspx?collectionid=390#1121964"
type="xSource">Angļu–latviešu–krievu informātikas termini</xref>
      <transacGrp>
        <transac type="transactionType">origination</transac>
        <transacNote type="responsibility">Extracted automatically by TaaS</transacNote>
        <date>2014-05-08</date>
      </transacGrp>
      <transacGrp>
        <transac type="transactionType">approval</transac>
        <transacNote type="responsibility">GornostayT</transacNote>
        <date>2014-05-16T14:50:42.018Z</date>
      </transacGrp>
      <admin type="status">approved</admin>
      <admin type="sourceIdentifier">http://www.eurotermbank.com/search.aspx?text=World%20Wide
%20Web&amp;langfrom=en&amp;langto=lv&amp;where=etb&amp;advanced=false#pos=1</admin>
      <transacGrp>
        <transac type="transactionType">modification</transac>
        <transacNote type="responsibility">GornostayT</transacNote>
        <date>2014-05-16T14:59:00.814Z</date>
      </transacGrp>
    </ntig>
    <note>Microsoft Public Terminology Collection</note>
  </langSet>
  <langSet xml:lang="en" xmlns:xml="http://www.w3.org/XML/1998/namespace">
```

```
<ntig>
<termGrp>
<term>World Wide Web</term>
<termNote type="MSD">NNP NNP NNP</termNote>
<termCompList type="lemma">
<termCompGrp>
<termComp>World</termComp>
<termNote type="partOfSpeech">noun</termNote>
<termNote type="grammaticalNumber">singular</termNote>
</termCompGrp>
<termCompGrp>
<termComp>Wide</termComp>
<termNote type="partOfSpeech">adjective</termNote>
<termNote type="grammaticalNumber">singular</termNote>
</termCompGrp>
<termCompGrp>
<termComp>Web</termComp>
<termNote type="partOfSpeech">noun</termNote>
<termNote type="grammaticalNumber">singular</termNote>
</termCompGrp>
</termCompList>
</termGrp>
<descripGrp>
<descrip type="surface">World Wide Web</descrip>
<descripNote type="surfaceForm">World Wide Web</descripNote>
<descripNote type="surfaceMSD">NNP NNP NNP</descripNote>
<descripNote type="normalizedForm">World Wide Web</descripNote>
<descripNote type="normalizedMSD">NNP NNP NNP</descripNote>
<descripNote type="lemma">World Wide Web</descripNote>
<descripNote type="context">nature of the<hi>World Wide Web</hi>for communication
in</descripNote>
<descripNote type="file">About The Project.htm</descripNote>
</descripGrp>
<descripGrp>
<descrip type="surface">World Wide Web</descrip>
<descripNote type="surfaceForm">World Wide Web</descripNote>
<descripNote type="surfaceMSD">NNP NNP NNP</descripNote>
<descripNote type="normalizedForm">World Wide Web</descripNote>
<descripNote type="normalizedMSD">NNP NNP NNP</descripNote>
<descripNote type="lemma">World Wide Web</descripNote>
<descripNote type="context">viability of the<hi>World Wide Web</hi>is paramount.
In</descripNote>
<descripNote type="file">About The Project.htm</descripNote>
</descripGrp>
<descripGrp>
<descrip type="surface">World Wide Web</descrip>
<descripNote type="surfaceForm">World Wide Web</descripNote>
<descripNote type="surfaceMSD">NNP NNP NNP</descripNote>
<descripNote type="normalizedForm">World Wide Web</descripNote>
<descripNote type="normalizedMSD">NNP NNP NNP</descripNote>
<descripNote type="lemma">World Wide Web</descripNote>
<descripNote type="context">driven by the<hi>World Wide Web</hi>Consortium (W3C),
an</descripNote>
<descripNote type="file">About The Project.htm</descripNote>
</descripGrp>
<transacGrp>
<transac type="transactionType">origination</transac>
<transacNote type="responsibility">Extracted automatically by TaaS</transacNote>
<date>2014-05-08</date>
</transacGrp>
<xref target="http://taas.eurotermbank.com/extractor" type="xSource" />
<descripGrp>
```

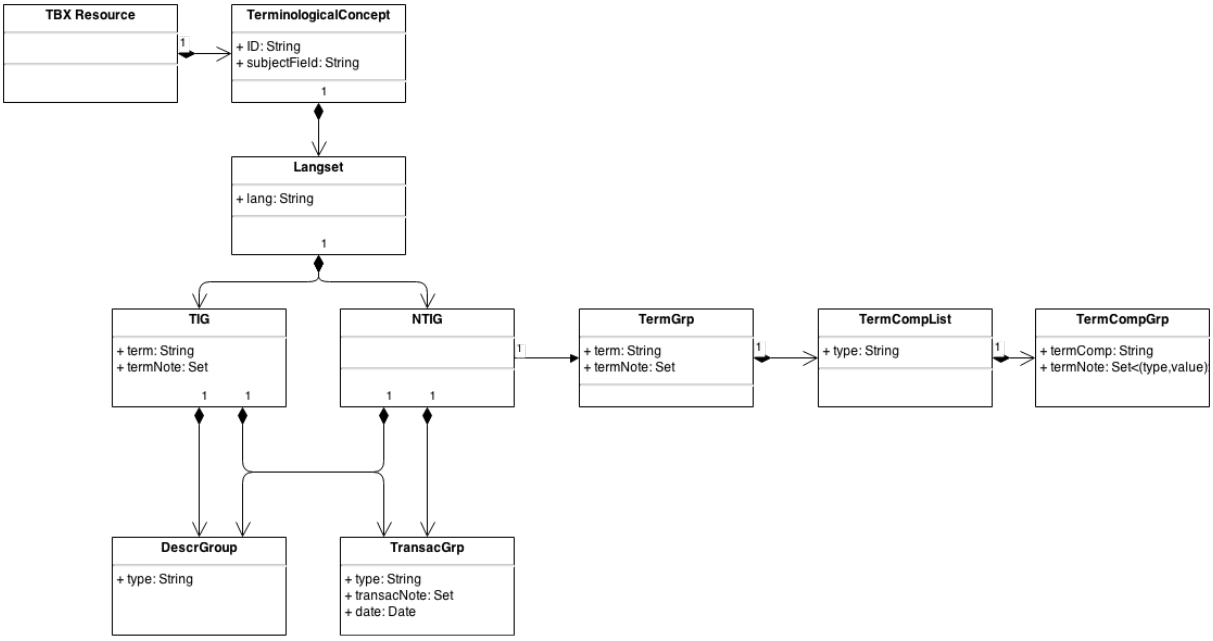
```
      <descrip type="context">nature of the<hi>World Wide Web</hi>for communication in</descrip>
      <admin target="About The Project.htm" type="sourceIdentifier" />
    </descripGrp>
    <descripGrp>
      <descrip type="context">viability of the<hi>World Wide Web</hi>is paramount. In</descrip>
      <admin target="About The Project.htm" type="sourceIdentifier" />
    </descripGrp>
    <descripGrp>
      <descrip type="context">driven by the<hi>World Wide Web</hi>Consortium (W3C), an</descrip>
      <admin target="About The Project.htm" type="sourceIdentifier" />
    </descripGrp>
    <transacGrp>
      <transac type="transactionType">approval</transac>
      <transacNote type="responsibility">GornostayT</transacNote>
      <date>2014-05-16T14:50:08.467Z</date>
    </transacGrp>
    <admin type="status">approved</admin>
    <admin type="sourceIdentifier">http://www.eurotermbank.com/search.aspx?text=World%20Wide
%20Web&amp;langfrom=en&amp;langto=lv&amp;where=etb&amp;advanced=false#pos=1</admin>
    <transacGrp>
      <transac type="transactionType">modification</transac>
      <transacNote type="responsibility">GornostayT</transacNote>
      <date>2014-05-16T14:59:00.821Z</date>
    </transacGrp>
  </ntig>
  <note>Microsoft Public Terminology Collection</note>
  <descrip type="definition">A set of interlinked documents in a hypertext system. The user enters the
web through a home page.</descrip>
 </langSet>
</termEntry></body>
 </text>
</martif>
```

Basic TBX Data Model

The following figure summarizes the TBX Data Model as an UML diagram. We briefly describe the TBX Data Model abstracting from the XML specifics in what follows:

- **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:** represents a language-independent concept. Each terminological concept is associated to a *LangSet*, 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).
- **TIG:** represents a language-specific term for which no decomposition information is provided.

- **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.



Mapping TBX to RDF

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**.

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 a Bitbucket project: https://vroddon@bitbucket.org/vroddon/tbx2rdf.git
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.

Output

TO BE INSERTED

Best Practices and Recommendations