TBX in ODD: Schema-agnostic specification and documentation for TermBase eXchange

Stefan Pernes
INRIA
stefan.pernes@inria.fr

Laurent Romary
INRIA
laurent.romary@inria.fr

Kara Warburton
Termologic
kara@termologic.com

Abstract

TermBase eXchange (TBX), the ISO standard for the representation and interchange of terminological data, is currently undergoing revision and will for the first time formalize overarching structural constraints regarding the definition and validation of dialects and XML styles. The paper describes the design of an ODD architecture, which allows for a complete specification of present-day TBX.

1 Introduction

TermBase eXchange (TBX), the ISO standard for the representation and interchange of terminological data, is currently undergoing revision and will for the first time formalize overarching structural constraints regarding the definition and validation of dialects and XML styles. To match these requirements, the ODD specification language provides advanced subset selection and constraint specification capabilities covering both structure and content of text encoding formalisms. Following the literate programming methodology, it furthermore allows the definition of integrated resources, which contain formal specifications alongside their prose descriptions and usage examples. This paper first describes the meta-model behind TBX as well as current challenges in the context of its revision. From this follows a description of applicable ODD mechanisms and the design of an ODD architecture, which allows for a complete specification of present-day TBX.

2 Terminological Markup Framework and TermBase eXchange

Terminology standards have witnessed a long evolution since 1987 when a first pre-SGML format for storage and interchange via magnetic tapes was devised. Published in 2003 as ISO 16642 (2003), the Terminological Markup Framework (TMF) marks a pivotal point in the succession of terminology standards as it constitutes a meta structure for terminology encoding which also provides the foundation for the more recent TBX format. TMF *specifies a framework designed to provide guidance on the basic principles for representing data recorded in terminological data collections* (ibid.). Its aims can be described as twofold (Romary 2001, 2f): (1) as a meta-model for terminological data representation it facilitates the description and comparison of existing interchange formats, and (2) it provides a mechanism for the flexible definition of interchange formats while safeguarding interoperability between them. The specification of such a meta-model thus eases the integration of different terminological databases with each other as well as with other lexical resources. In principle, TMF allows one to describe a potentially infinite set of Terminological Markup Languages (TML). Formally, this flexibility is achieved by describing the various components of terminological databases as either part of the structural skeleton or as data categories. This leads to the four elementary notions of TMF (ibid., 3f):

1. The meta-model: A structural skeleton for terminological entries following a concept-oriented, or onomasiological, view.

2. Information units taken from a Data Category Repository (DCR) as described in ISO 12620 (1999) (a new version is about to be published).

3. Methods and representations: The actual implementation of a TML, combining the structural skeleton with the chosen data categories. This also comprises the mappings between data categories and the vocabularies used to express them (e.g. as an XML element or a database field).

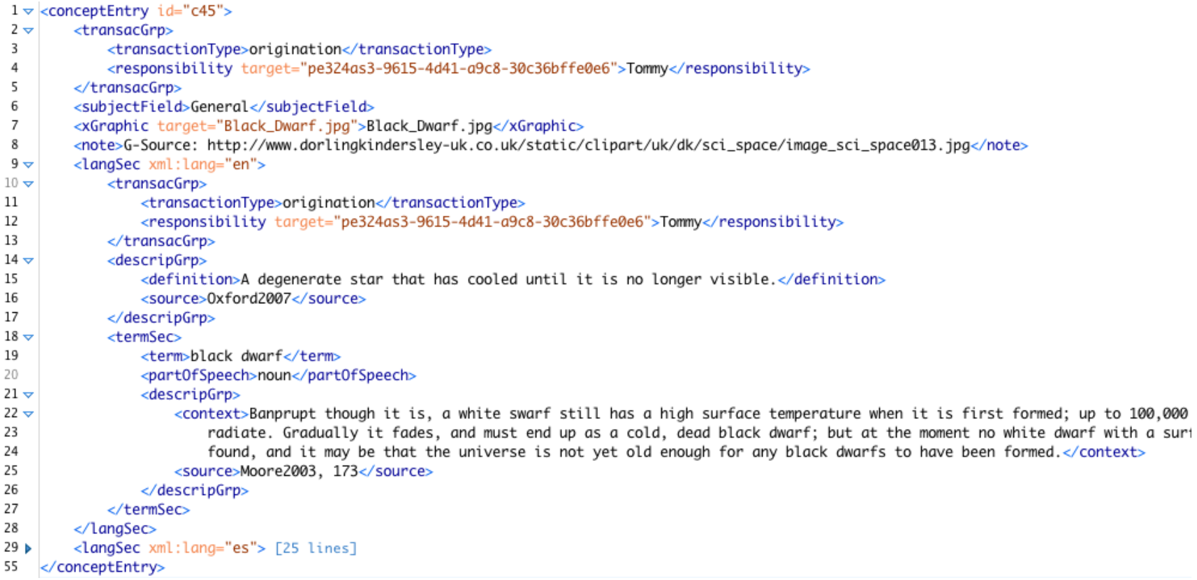
4. A generic mapping tool: A methodology that maps any given TML onto the meta-model. The notion of a generic mapping tool can be replaced by the ODD architecture as proposed in this paper.

As a consequence of these elementary notions, the interoperability between two TMLs is reduced to a comparison of their respective use of data categories.

TermBase eXchange (TBX) as defined by ISO 30042 (2008) is precisely an instantiation of the described meta-model alongside a specific selection of data categories. It is a reference implementation of TMF, taking the form of XML (thus its official specification is implemented as XML DTD, RelaxNG, and W3C schemas) and constituting what is nowadays called TBX-Default, or the master TBX dialect. It is *designed to support various types of processes involving terminological data, including analysis, descriptive representation, dissemination, and interchange (exchange), in various computer environments* (ibid.). TBX is currently undergoing extensive review as ISO CD30042 (2017): It establishes provisions for the specification of official TBX dialects, such as their minimum requirements in terms of structure and data categories. Furthermore, it is targeting increased interoperability by merging the content models of two alternative term information group elements, *tig* and *ntig*, into one *termSec* element. See Figure 1 for a short sample TBX entry (in accordance with the current Community Draft version). Additionally, two markup styles, data categories as tags (DCT) and data categories as attribute values (DCA), are defined. Apart from the need for a modular framework for dialect specification, a consequence of this development is that the specification of TBX itself will profit from tighter control over data categories, their identifiers and values, as well as data type information for these values. As has also been noted in the context of data exchange with Linked Open Data description formalisms such as RDF and OWL, the current specification of TBX does not provide data type description mechanisms in the classical sense, focussing mostly on string values based on W3C XML primitives (cf. Reineke 2014, 7). Thus, the strengths of an ODD architecture as proposed herein are a modular mechanism for the description of TBX dialects and an increased control over data types.

3 The Text Encoding Initiative and *One Document Does it all*

The Text Encoding Initiative (TEI) maintains a set of guidelines which have become a de facto standard in the encoding of literary, historical, and linguistics research data. Being established in 1987, the TEI guidelines predate and inform a number of modern web and encoding standards. They comprise close to 500 elements which are organised in functional-thematic modules, classes of shared attributes, and macros for common content models. The broadness and complexity of the TEI is paired with its own specification language that allows for a modular definition of project-specific customizations. One Document Does it all (ODD) is a generic specification language, establishing a separation between the specification of TEI encoding models and current schema languages, be it a XML DTD, a RelaxNG schema, or a W3C schema. Thus, TEI encoding models are essentially agnostic about the choice of a representation language and could also map to formalisms other than XML (Burnard 2013, 13). Furthermore, ODD follows the literate programming paradigm and constitutes a single resource containing formal declarations alongside descriptive prose and examples of usage (Burnard and Rahtz 2004, 3). Figure 2 schematically displays the different aspects of processing ODD files.



```
<conceptEntry id="c45">
    <transacGrp>
        <transactionType>origination</transactionType>
        <responsibility target="pe324as3-9615-4d41-a9c8-30c36bffe0e6">Tommy</responsibility>
    </transacGrp>
    <subjectField>General</subjectField>
    <xGraphic target="Black_Dwarf.jpg">Black_Dwarf.jpg</xGraphic>
    <note>G-Source: http://www.dorlingkindersley-uk.co.uk/static/clipart/uk/dk/sci_space/image_sci_space013.jpg</note>
    <langSec xml:lang="en">
        <transacGrp>
            <transactionType>origination</transactionType>
            <responsibility target="pe324as3-9615-4d41-a9c8-30c36bffe0e6">Tommy</responsibility>
        </transacGrp>
        <descripGrp>
            <definition>A degenerate star that has cooled until it is no longer visible.</definition>
            <source>Oxford2007</source>
        </descripGrp>
        <termSec>
            <term>black dwarf</term>
            <partOfSpeech>noun</partOfSpeech>
            <descripGrp>
                <context>Banprupt though it is, a white swarf still has a high surface temperature when it is first formed; up to 100,000
                    radiate. Gradually it fades, and must end up as a cold, dead black dwarf; but at the moment no white dwarf with a sur
                    found, and it may be that the universe is not yet old enough for any black dwarfs to have been formed.</context>
                <source>Moore2003, 173</source>
            </descripGrp>
        </termSec>
    </langSec>
    <langSec xml:lang="es"> [25 lines]
</conceptEntry>
```

Figure 1: Example TBX entry in DCT style

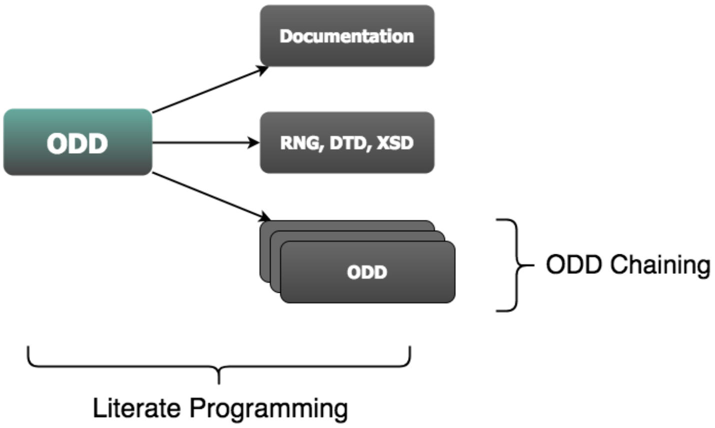
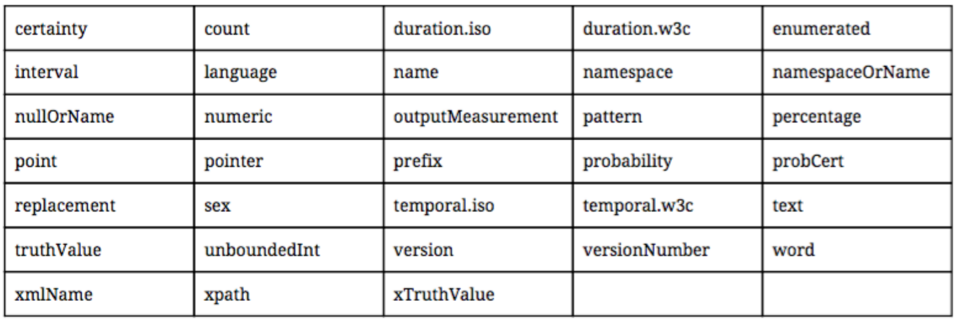


Figure 2: ODD processing and chaining



certainty	count	duration.iso	duration.w3c	enumerated
interval	language	name	namespace	namespaceOrName
nullOrName	numeric	outputMeasurement	pattern	percentage
point	pointer	prefix	probability	probCert
replacement	sex	temporal.iso	temporal.w3c	text
truthValue	unboundedInt	version	versionNumber	word
xmlName	xpath	xTruthValue		

Figure 3: TEI datatypes

Formal declarations in ODD concern foremost the key components of the TEI abstract model: elements, attributes, modules, classes, and macros. The last three components serve to reduce the overall systems complexity and allow a coarse-grained selection of characteristics for a TEI customization. Another major simplification stems from a conscious effort to provide uniform levels of description and hence processing allowing components to be added, changed, replaced, or deleted within a given context and at any point in a schema declaration (Burnard and Rahtz 2004, 8). Modifications of this kind can be chained together (ODD chaining), thus making it easy to supplement a broadly specified TEI customization with fine-grained, context-specific modifications. The context specificity of such declarations also allows for a tighter constraint on possible attribute values by means of data typing (Burnard 2013, 10): The vast majority of attribute values are defined by reference to a data type macro as defined within the ODD system, which are in turn mapped to a W3C Schema data type or to an expression in RELAX NG syntax, thus allowing the ODD system to overlay additional semantics onto such bare data types (see Figure 3, TEI datatypes). Additionally, a further layer of constraint specification can be added using ISO Schematron, making it possible to implement many of the informally expressed rules for good practice, which are typically found in the prose of encoding guidelines.

4 Description of the *TBX in ODD* architecture

Prior work has already established an ODD architecture for the ISO 30042 (2008) *Basic* dialect (Romary 2014). Furthermore, the TBX specification published in 2008 was actually written in ODD by the core editorial team of the time[1]. This approach – a major diversion from conventional ISO authoring practices – was undertaken as case study for the use of ODD to author ISO standards that contain a mixture of prose, machine-readable specifications and sample code, and that require schemas as derivative products. Major changes introduced by the current revision of TBX are a modular framework for the specification of dialects and the implementation of markup styles (DCA vs. DCT). Both requirements can be met via module selection in ODD. A prerequisite for this ODD chaining mechanism is a master ODD file, containing specifications for all core structure elements and permissible data categories (it is equivalent to TBX-Default, the so-called master TBX dialect). Specifying a new TBX dialect in ODD is achieved by selecting modules and classes, or parts thereof, and selecting data categories from the ISO Data Category Repository, which may also extend on the TBX default set of data categories. Using the ODD framework, this dialect can afterwards be transformed into any of the supported schema languages for document validation – which also serve as the official specification of TBX and its dialects. The same approach applies for the validation of one or the other data category style. For example, in order to generate a schema for validating files using only DCA style, one would define a subset of the master ODD that excludes the module *TBXDCT* and vice versa.

A preliminary overview of the module and class organization is shown in Figure 4. File header elements, core structure, and elements specific to the two markup styles are grouped into modules, while

[1] Arle Lommel, Alan Melby, and Kara Warburton

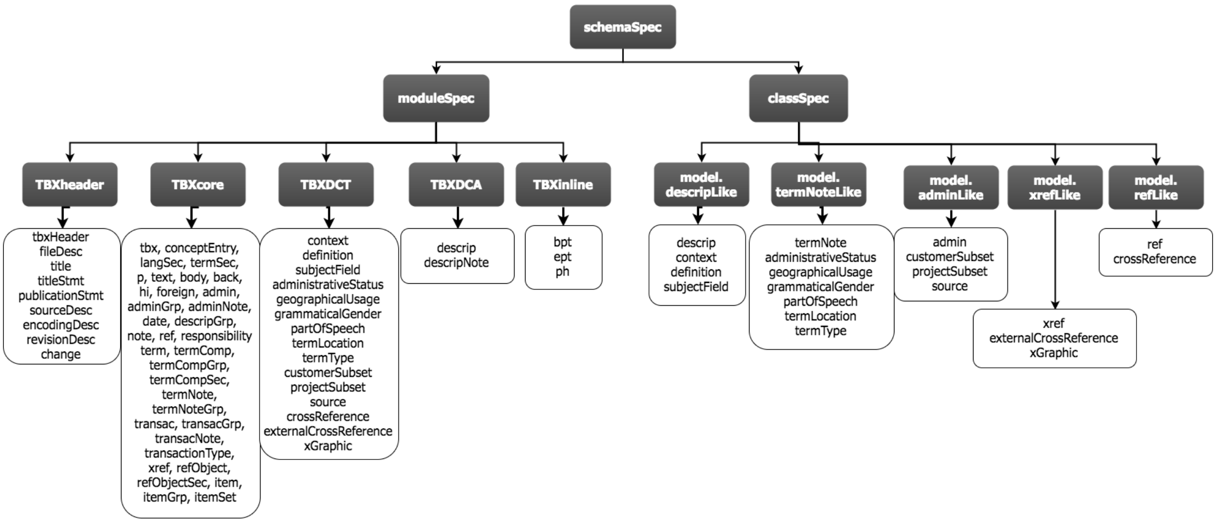

schemaSpec
moduleSpec
classSpec
TBXheader
TBXcore
TBXDCT
TBXDCA
TBXinline
model. descripLike
model. termNoteLike
model. adminLike
model. xrefLike
model. refLike
tbxHeader fileDesc title titleStmt publicationStmt sourceDesc encodingDesc revisionDesc change
tbx, conceptEntry, langSec, termSec, p, text, body, back, hi, foreign, admin, adminGrp, adminNote, date, descripGrp, note, ref, responsibility term, termComp, termCompGrp, termCompSec, termNote, termNoteGrp, transac, transacGrp, transacNote, transactionType, xref, refObject, refObjectSec, item, itemGrp, itemSet
context definition subjectField administrativeStatus geographicalUsage grammaticalGender partOfSpeech termLocation termType customerSubset projectSubset source crossReference externalCrossReference xGraphic
descrip descripNote
bpt ept ph
descrip context definition subjectField
termNote administrativeStatus geographicalUsage grammaticalGender partOfSpeech termLocation termType
admin customerSubset projectSubset source
xref externalCrossReference xGraphic
ref crossReference

Figure 4: TBX module and class organization in ODD

elements with similar content models are grouped into classes. The interlocking nature of the organization hints at the powerful subset selection mechanism. As the revision of TBX is still ongoing, some data category and constraint specifications are yet to be implemented.

5 Conclusion

In this paper we have described the specification of TermBase eXchange (TBX), which is currently undergoing revision as ISO CD30042 (2017), using the ODD specification language. The requirements for this upcoming version of TBX include a system for the derivation of dialects, which need to be verifiably compliant to the core structure and, in the best case, can be defined in a user-friendly, modular fashion. The ODD language provides such a framework and follows a literate programming approach where documentation, usage examples, and formal specifications all reside in one document. Additionally, it is a sustainable approach that does not depend on any specific schema language and is in principle able to map to the data modelling ecosystem of the day – an advantage given the long-term perspective of terminology encoding standards.

References

Burnard, L. (2013). Resolving the Durand Conundrum. *Journal of the Text Encoding Initiative 6*. `http://jtei.revues.org/842`.

Burnard, L. and S. Rahtz (2004). RelaxNG with son of ODD. *Extreme Markup Languages*. `http://www.tei-c.org/cms/Talks/extreme2004/paper.html`.

ISO 12620 (1999). Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources. Standard, International Organization for Standardization, Geneva, Switzerland.

ISO 16642 (2003). Computer applications in terminology – Terminological markup framework. Standard, International Organization for Standardization, Geneva, Switzerland.

ISO 30042 (2008). Systems to manage terminology, knowledge and content – TermBase eXchange (TBX). Standard, International Organization for Standardization, Geneva, Switzerland.

ISO CD30042 (2017). Systems to manage terminology, knowledge and content – TermBase eXchange (TBX). Standard, International Organization for Standardization, Geneva, Switzerland.

Reineke, D. (2014). TBX between termbases and ontologies. *Proceedings of the 11th International Conference on Terminology and Knowledge Engineering, TKE 2014.* `http://hal.archives-ouvertes.fr/hal-01005838`.

Romary, L. (2001). An abstract model for the representation of multilingual terminological data: TMF – Terminological Markup Framework. *Proceedings of the 5th TermNet Symposium, TAMA 2001.* `http://hal.inria.fr/inria-00100405`.

Romary, L. (2014). TBX goes TEI – Implementing a TBX basic extension for the Text Encoding Initiative guidelines. *Proceedings of the 11th International Conference on Terminology and Knowledge Engineering, TKE 2014.* `hal-00950862v2`.