Journal of the Text Encoding Initiative

Issue 3 (November 2012)
TEI and Linguistics

Gerhard Budin, Stefan Majewski and Karlheinz Mörth

# Creating Lexical Resources in TEI P5

## A Schema for Multi-purpose Digital Dictionaries



Electronic reference
Gerhard Budin, Stefan Majewski and Karlheinz Mörth, « Creating Lexical Resources in TEI P5 », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 05 November 2012. URL : http://jtei.revues.org/522 ; DOI : 10.4000/jtei.522

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on:
http://jtei.revues.org/522
Document automatically generated on 05 November 2012.

**Gerhard Budin, Stefan Majewski and Karlheinz Mörth**

# Creating Lexical Resources in TEI P5

A Schema for Multi-purpose Digital Dictionaries

## 1. Background

1 Lexicography, the art of compiling dictionaries, is one of the oldest branches of linguistics. All remnants of early lexicographic writings stem from Asia, and the oldest extant precursors of modern dictionaries were Sumerian/Akkadian clay tablets dating from the second millennium BC; these early lexicographic endeavours represent a very modern type of text—a bilingual dictionary (Snell-Hornby 1986, 208) which, in most areas of the world, would not emerge until at least 2,000 years later.

2 In contrast to the Sumerian clay tablets, most other early testimonies of this academic tradition were monolingual in nature. The Sanskrit grammarian *Yāska*[1] is regarded by many as the earliest known Indian lexicographer; his Nirukta was a treatise on etymology and semantics, containing a glossary of irregular verbs. Chinese lexicography is some centuries younger: the *Erya* (author unknown) is the most ancient Chinese writing that falls into the broader category of dictionaries (Wilkinson 2000, 62).

3 Although the creation of modern dictionaries is considered to have begun in Europe with the rise of national languages, there is no clearly discernible demarcation line between pre-modern and modern dictionary production. Some outstanding works emerged in the 17th and 18th centuries. Jean Nicot's *Trésor de la langue Française* was printed in 1606, Agnolo Monosini's *Vocabulario della lingua italiana* appeared in 1612, Johann Christoph Adelung's *Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart* followed in 1781, and Samuel Johnson finished his *Dictionary of the English Language* in 1755.[2] The first large-scale Chinese dictionary from this time period, the *Kangxi zidian*, dates from 1716 (Wilkinson 2000, 64).

4 The latest step in this long history is being constituted by the transition towards digital methods. Today, digital technology is not only used to produce print dictionaries; rather, many dictionaries exist solely in digital form. Information and communication technology has become pervasive in all stages of the modern dictionary creation process: both data acquisition and representation of lexical knowledge rely heavily on this technology. Furthermore, dictionary makers have shifted from traditional methods such as introspection and interviews of competent speakers towards more empirical methods based on lexicographic research using increasingly sophisticated digital resources such as corpora (large digital text collections that reflect real-world language usage).

## 2. The ICLTT's Dictionaries

5 The Institute for Corpus Linguistics and Text Technology (ICLTT) of the Austrian Academy of Sciences has been conducting a number of lexicographic projects, including both digitizing print dictionaries and creating born-digital lexicographic data. The lexicographic data produced in these projects are designed to serve a variety of purposes for both linguistic research and lexicography. To ensure that NLP tools available at the institute would work with all the data, a uniform encoding system for all projects was needed. The integration of digital corpus data with the lexicographic infrastructure has been an important goal and plays an important role in all these efforts.

6 The ICLTT as an institution has grown out of several projects. One of the best known results of these projects is probably the Austrian Academy Corpus (AAC), a digital collection of German language texts stemming from the 19th and 20th centuries. The digital texts contained in the AAC were collected with a literary, a socio-historic and a lexicographic perspective in mind, but in spite of the literary and historical focus in setting up the corpus, it is increasingly used by linguists (Moerth 2002).

## 2.1. Print Dictionaries

7 The main motive behind setting up the corpus was the institute's involvement in a longstanding text-lexicographic project which produced two dictionaries designed to ease access to one of Austria's most important works of twentieth-century literature, Karl Kraus' magazine *Die Fackel*. The first volume was a dictionary of idioms and idiomatic expressions; the second one a comprehensive listing and documentation of insults and invective terms.

8 In recent years, the institute has shifted from addressing the needs of literary scholars by focusing on particular works of literature to catering to the needs of linguists by devoting resources to smaller and more diverse projects. The ICLTT has also contributed to the production of the largest German-Russian dictionary ever produced (Dobrovolsky 2008–2010), which was published as a cooperative project of the Austrian and the Russian Academies of Sciences.

9 In addition to creating new print dictionaries, the institute has also digitized historical dictionaries and even incorporated them into the AAC in order to extend the collection of texts to as many types of written language as possible. Currently, efforts are being made to make this data TEI P5 compliant.

## 2.2. Born-digital Dictionaries

10 Dictionaries are increasingly created in and for the digital world. Apart from digitizing paper dictionaries, the ICLTT has also started to create new digital lexical resources, some of which build on the department's digital text collections. These include dictionaries for doing variational linguistics on German as written and spoken in Austria, Early Modern German, and Arabic; a GUI tool for converting German Wiktionary data to TEI P5;[3] and a comprehensive *Dictionary of Modern Persian Single Word Verbs* to be used as the basis for a morphological analyzer. The variation among these projects has been brought about to a certain degree by the ICLTT's role as Austria's CLARIN and DARIAH coordinator.

# 3. Data Formats

11 In choosing a uniform encoding system for all ICLTT data, the department's staff surveyed data formats in use. Although most of the relevant dictionary productions of the recent past have relied on digital data and methods, there is little consensus on standards. A great number of divergent formats have coexisted: MULTILEX and GENELEX (GENEric LEXicon) are systems that are associated with the Expert Advisory Group on Language Engineering Standards (EAGLES).[4] Other formats used in digital dictionary projects are OLIF (Open Lexicon Interchange Format),[5] MILE (Multilingual ISLE Lexical Entry),[6] LIFT (Lexicon Interchange Format),[7] OWL (Web Ontology Language)[8] and DICT (Dictionary Server Protocol),[9] the latter being an important dictionary delivery format (Faith 1997).

12 Another standard considered was ISO 1951 ("Presentation/representation of entries in dictionaries – requirements, recommendations and information"). Although this standard focuses on encoding the presentation of lexicographical data in dictionaries for human use in what is called LEXml (Lexicographical Markup Language), it seems that after a few years of existence only few publishing houses have been using this format (such as Langenscheidt, Munich) for their dictionary production line.

13 Last but not least, when looking for an encoding standard for machine readable dictionaries, ISO 24613:2008 ("Language resource management – Lexical markup framework (LMF)"), the ISO standard for natural language processing (NLP) and machine-readable dictionaries (MRD), must be considered. Recently, there have been discussions about the possibility of creating a TEI serialization of LMF (Romary 2010).

14 In modeling lexicographic data, it has become common practice to conceptualize the underlying structures as tree-like constructs, which makes XML an ideal syntax for expressing the data. Another option, from software engineering, is UML (Unified Modeling Language)[10] which in turn can easily be serialized into an XML vocabulary. This approach was taken by the authors of LMF.

15 For our projects, the final "short list" contained ISO 1951, LMF and the TEI dictionary module. ISO 1951 was eschewed from the very beginning, among other reasons for lack of support in the community. LMF in turn has gained more support in the dictionary-producing community. Given the still small amount of available data using LMF and ongoing discussions, the decision was made to move towards TEI and keep an eye on the LMF specification as it develops.[11]

## 4. TEI Dictionary Module

16 The TEI dictionary module appears to be the de facto encoding standard for dictionaries digitized from print sources. As such, "TEI for dictionaries" has a longstanding tradition. Interestingly, the most recent versions of the TEI Guidelines contain a passage that indicates that the authors had in mind a much wider range of dictionaries:

> ... The elements described here may also be useful in the encoding of computational lexica and similar resources intended for use by language-processing software; they may also be used to provide a rich encoding for word lists, lexica, glossaries, etc. included within other documents. (TEI Consortium P5 2012, 247)

17 This passage reflects a considerable conceptual extension of the initial purpose of the module.[12] However, the idea of extending the scope of the TEI dictionary module for use by language-processing software is not at all as far-fetched as it may seem at first glance. The fact that there are people interested in the issue has been documented by the large audience of the workshop "Tightening the Representation of Lexical Data: A TEI Perspective," held at the 2011 Annual Conference and Members' Meeting of the TEI Consortium (Würzburg, Germany). Actually, the TEI's ability to adapt to many types of dictionaries makes it an ideal candidate for such an endeavor.

18 A fundamental problem we came up against when we started to model our dictionary data was the lack of available examples against which we could compare our data. It would have been beneficial if more projects had made at least samples of their data publicly accessible.[13] Many of the examples which can be found on the TEI website are repetitive and are by no means exhaustive.[14] However, getting hold of examples in other encoding languages is not easy either: ISO 1951 seems to be used by a single publishing house and LMF has not won much ground in the field, though there are some data available for the latter.[15]

## 5. ICLTT's TEI Schema

19 The following sections outline selected features of the ICLTT's customization of the TEI P5 dictionary module. The system has been used successfully for lexicographic data encoding at the department, where it is meant to be a multi-purpose system targeting both human users and software applications. The following four requirements had featured strongly in our decision in favor of TEI encoding:

- Acquaintance with the overall TEI system: as the department has been working with TEI on text encoding projects, a number of colleagues are conversant with TEI and have used it from the very beginning of our dictionary projects;
- Intuitiveness of the TEI system: the concise and yet expressive set of elements is definitely more easily readable to human lexicographers working on the XML source than for instance the LMF serialization proposed in ISO 24613:2008;
- Consistency with other language resources contained in the same collection: the intention was to keep the encoding system of the dictionary resources in line with other textual data to be integrated with these lexicographic resources.
- Adaptability to the needs of dictionaries to be used in natural language processing (NLP).

20 In order to make the TEI dictionary module usable for NLP purposes, it has been necessary to tighten the many combinatorial options of TEI P5—that is, to constrain the content models of various elements.

## 5.1. Representing Lemmas

21 In TEI, dictionaries are a specific type of text and are therefore encoded with <text> elements, which are made up of optional <front> and <back> matter. The dictionary entries are placed in a <body> element.

```
<TEI>
   <teiHeader>
   ...
   </teiHeader>
   <text>
      <front>...</front>
      <body>
         <entry>...</entry>
         <entry>...</entry>
         <entry>...</entry>
         ...
         ...
         ...
      </body>
      <back>...</back>
   </text>
</TEI>
```

22 Individual entries may be seen as the core of all lexicographic encoding; the structure of dictionary entries can display a great variety of different forms.[16] This also accounts for the fact that the P5 version of the Guidelines (250) offer three elements to encode this type of microtexts: <entry>, <entryFree>, and <superEntry>.

23 The <superEntry> element can be used to group entries together and is not used in our schema. As the name implies, <entryFree> contains a single <entry> with a comparatively large number of acceptable elements that may be arranged in many different ways. In TEI P5, <entryFree> can contain 30 different elements from the dictionary module alone.[17] The great flexibility of this element makes it suitable for digitizing print dictionaries, but in creating strictly defined dictionary structures to be used by software, this flexibility is of lesser value.

24 In contrast to <entryFree>, the <entry> element allows for only ten sub-elements: <case>, <def>, <etym>, <form>, <gramGrp>, <hom>, <sense>, <usg>, <xr>, and <dictScrap>. The dictionary schema described in this paper only contains the simple <entry> element (combinatorial options were further restricted by excluding both <dictScrap> and <hom> elements from the list of possible child elements).

25 Simple dictionary entries invariably start with a lemma. Optionally, entries contain an indication of the word class of the lemma and one or more <sense> elements. A typical entry has a structure like this:

```
   <entry>
      <form type="lemma">
      ...
      </form>
      <gramGrp>
         <gram type="pos">...</gram>
      </gramGrp>
      <sense>
      ...
      </sense>
...</entry>
```

26 In many cases, it is difficult for lexicographers to decide whether to integrate lexical items into one single entry or rather to make two or more entries. Lexical homonymy in TEI dictionaries is often encoded using the <hom> element, as in the following abridged example.

```
<entry>
   <form type="lemma"><orth>Schloss</orth></form>
   <hom>
      <sense>
         <cit type="translation" xml:lang="en">
```

```
            <quote>castle, palace</quote></cit>
      </sense>
   </hom>
   <hom>
      <sense>
         <cit type="translation" xml:lang="en">
            <quote>(pad)lock</quote></cit>
      </sense>
   </hom>
</entry>
```

27 As a basic principle, we have attempted to keep hierarchies in our encoding system as flat as possible. This is why the <hom> element has been excluded from the set of possible elements. That is, in cases of homonymy, lexicographers have to either work with entries that contain several senses or to create separate entries, which would be encoded in TEI as follows:

```
<entry>
   <form type="lemma"><orth>Schloss</orth></form>
   <sense>
      <cit type="translation" xml:lang="en">
         <quote>castle, palace</quote></cit>
      </sense>
</entry>
<entry>
   <form><orth>Schloss</orth></form>
   <sense>
      <cit type="translation" xml:lang="en">
         <quote>(pad)lock</quote></cit>
      </sense>
</entry>
```

28 The same encoding pattern is applied to grammatical homonyms and polyfunctional items—that is, homographs that are semantically related but have different *word classes*. However, encoding homonyms in separate <entry> elements can be problematic, especially when lexical items belong to different word classes and need to be distinguished (consider an example from English: “talk” as a verb versus as a noun). For us, the deciding factor was whether the word class difference manifests itself in the semantic description, the <sense> block in TEI nomenclature. Whenever different part-of-speech labels would need to be assigned to <sense> elements (such as with all grammatical homonyms), the lexical items were encoded in separate <entry> elements rather than in one.

29 Polyfunctionality is a very common phenomenon and has posed problems in almost all our projects. Our approach, as detailed above, has pros and cons. However, our main argument in favor of splitting entries—putting each homonym into a separate <entry>—is that it makes access to the particular lexical items more straightforward. Working along these lines, part-of-speech labels only appear on the top-most level of the entry together with the lemma, not within <sense> elements. If necessary, the relation between entries could be made explicit by <re> (related entry) elements or some system of links.

30 It is obvious that the decision of whether to split entries also depends on what one plans to do with a particular set of data. For some of our projects, we have plans to enrich lexical data using corpora: looking for new, hitherto unregistered word forms, doing statistics on word forms, etc.

## 5.2. Encoding Word Class Information

31 A fixed component of all single-word dictionary entries is a block containing word-class information. In early experiments, we encoded this information within the <form> element representing the lemma. While TEI allows word-class information to appear in various locations within an <entry> element, the motivation behind putting it within <form> was that it seemed to be more consistent to say that the lemma, rather than the entry, belongs to a particular word class. In addition, putting the <gramGrp> element in the lemma’s <form> element allowed <gramGrp> elements containing part-of-speech information to appear inside <form> elements, yielding an additional simplification of the schema.

32 Over time, we have come back to a more canonical TEI encoding, abandoning this rather atypical practice. This change of attitude was, among other things, motivated by experiments of converting our data into an LMF-conformant XML serialization: in LMF, @part-of-speech is defined as an attribute of the element <LexicalEntry>.[18]

33 Practical experience has also led us to change usage of elements inside the <gramGrp> element. Initially, word-class information was encoded using the <gramGrp> element, which can contain a number of other elements such as <case>, <gen>, <mood>, <pos>, and <tns>. For example:

```
...
<gramGrp>
   <pos>noun</pos>
</gramGrp>
...
```

34 We now only allow the <gram> element within <gramGrp>, using attributes to distinguish various word-class categories. The above example can be rewritten to its <gram> equivalent like this:

```
...
<gramGrp>
   <gram type="pos">noun</gram>
</gramGrp>
...
```

35 Choice of appropriate terminology is important when labeling lemmas with word classes. Scholars working on digital resources have long needed to maintain consistency both within a project and one agreed upon by the community at large. Nowadays, it also involves interoperability with other digital resources, especially by referring to publicly accessible frameworks (concept repositories) to make the linguistic terminology explicit. In the field of linguistics, two such frameworks play an increasingly important role: the so-called GOLD Standard, the General Ontology of Linguistics Descriptions (Farrar and Langendoen 2003) and ISOcat, the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al. 2009). The most important feature of the web-based ISOcat registry is that it provides persistent identifiers (PIDs) for all the concepts registered in the database, allowing for explicit reference to terms used.

36 So far, we have attempted to make use of ISOcat terminology in the ICLTT customization without explicitly referring to the ISOcat terms in the encoding of the entries. However, we have started to experiment with an alternative way of marking up word-class information that makes explicit reference to the concept repository which is exemplified in the following excerpts:

```
...
<gramGrp>
   <gram type="pos" corresp="#vrbNoun"/>
</gramGrp>
...
```

37 The label of the @corresp attribute above refers to a feature structure that, in turn, provides an explicit reference to the particular entry in the ISOcat database:

```
<fs type="partOfSpeech">
   <f xml: name="verbalNoun" fVal="http://www.isocat.org/datcat/DC-3858"/>
   <f xml: name="commonNoun" fVal="http://www.isocat.org/datcat/DC-385"/>
   <f xml: name="properNoun" fVal="http://www.isocat.org/datcat/DC-384"/>
</fs>
```

## 5.3. Morphosyntactic Information

38 Dictionary entries often contain more grammatical forms of the headword. In traditional lexicography, particular word forms are usually given in order to point the user to irregularities

in inflectional paradigms. In a digital dictionary, which does not have any spatial limitations, it is not uncommon to have more comprehensive lists of word forms.

### *5.3.1. <gramGrp> vs. Feature Structures*

39 The ICLTT has experimented with entries giving only inflectional irregularities and also those giving complete paradigms; in either case, each word form is encoded with a <form> element. Whatever the intended use of these word forms, a system is needed to identify their function. The traditional TEI way to do this would be to enter the morphosyntactic details of a <form> in a <gramGrp> element:

```
...
   <form type="inflected">
      <gramGrp>
         <pos value="verb"/>
         <tns value="present"/>
         <number value="singular"/>
         <mood value="indicative"/>
         <per value="2"/>
      </gramGrp>
      <orth>gehst</orth>
   </form>
...
```

40 In search of a more generic approach, we resorted to a system combining feature structures[19] and ISOcat grounded values. Instead of using the <gramGrp> element as a child of <form>, the @ana (analytic) attribute is added to the <form> element.

```
...
   <form type="inflected" ana="#v_pres_ind_sg_p2">
      <orth>gehst</orth>
   </form>
...
```

41 The labels used to construct the pointers in the @ana attribute are human-readable abbreviations. In this part of the system, we have attempted to proceed in line with the ISO TC37/SC4–related MAF (Morphosyntactic Annotation Framework) draft specification, in particular Chapter 8 on morpho-syntactic content (ISO 24611 2008, 21). The components of the value of the @ana attribute are resolved in a feature structure library:

```
<fvLib>
   ...
   <fs xml: name="v_pres_ind_sg_p2"
                feats="#pos.verb #tns.pres #mood.ind #num.pl #pers.2">
   ...
</fvLib>
<fLib>
   <f xml: name="pos"><symbol value="verb"/></f>
   ...
   <f xml: name="tense"><symbol value="present"/></f>
   ...
   <f xml: name="mood"><symbol value="indicative"/></f>
   ...
   <f xml: name="number"><symbol value="plural"/></f>
   ...
   <f xml: name="person"><symbol value="2nd"/></f>
   ...
</fLib>
```

42 This method of annotating morphosyntactic phenomena is not only extremely concise (the information is only referenced through links), it also allows for the assignment of multiple interpretations of the content of the <orth> element. The attribute @ana can contain an open number of so-called data.pointers, each separated by whitespace:

```
...
   <form type="inflected" ana="#v_pres_ind_pl_p1 #v_pres_ind_pl_p3 ">
      <orth>gehen</orth>
   </form>
```

```
...
```

### 5.3.2. A Particular Case: Encoding Roots of Semitic Words

43 Any general-purpose system such as the TEI is bound to have conceptual gaps. A particular problem of our projects involving Semitic languages was how to deal with what in Semitic studies is commonly referred to as a *root*. In Semitic morphology, word forms are constructed on top of two, three, or four consonants. These consonants, which function as abstract linguistic units, form what is commonly called "the root", i.e. the semantic skeleton of all morphologically derived forms. The scholars working with and on the described encoding system were very reluctant to use the TEI element <form> for the particular purpose, as this would have meant stretching the semantics of the element too much. Roots are neither *word forms* nor *stems*. In order to avoid "tag abuse", we first experimented with the TEI's feature-structure capabilities. Here is an example taken from our *Colloquial Cairene Arabic Dictionary* (*safar* is Arabic for 'journey').

```
...
   <form type="lemma">
      <form type="lemma"><orth>safar</orth></form>
      <fs><f name="root"><string>sfr</string></f></fs>
...
```

44 However, our current practice is to encode the root of each lemma by means of the <gramGrp> element holding the word-class information. Adding an additional <gram> element to <gramGrp> appears to be a both concise and conceptually consistent solution to the problem:

```
...
<gramGrp>
   <gram type="pos">noun</gram>
   <gram type="root">sfr</gram>
</gramGrp>
...
```

## 5.4. Identifying Linguistic Varieties and Writing Systems

45 When encoding digital texts, linguistic varieties are usually identified using so-called language codes, of which there are several systems. An older (yet very versatile) system is Verbix Language Codes, which makes use of the old SIL codes.[20] LS-2010 (Linguasphere language codes) is a rather recent system which was published in 2000 and updated in 2010. It contains over 32,000 codes. The most widely used standard is ISO 639.

46 All these systems are incomplete and, if still being maintained, continue to evolve. A downside to all of them is the lack of support coming from the many scholarly disciplines involved in their use. In addition to the high (and ever changing) number of linguistic varieties on our globe, one additional aspect has to be taken into consideration: many linguists also need codes for historic linguistic varieties as well as for living varieties.

47 In TEI encoding, it has become common practice to make use of the global[21] attribute @xml:lang, incorporated into the TEI from the World Wide Web Consortium's XML Specification. TEI prescribes this attribute to identify both linguistic varieties and writing systems. In this hybrid approach, the value of the attribute should be constructed in accordance with *Best Current Practice 47* (BCP 47)[22] which in turn refers to and aggregates a number of ISO standards (639-1, 639-2, ISO 15924, ISO 3166).[23]

48 BCP 47 defines an extensible system that is sufficiently expressive to identify most standard linguistic varieties. Language tags are assembled from a sequence of components (which are also called subtags), each separated by a hyphen. All subtags except for the first one are optional and have to be arranged in a particular order. The first subtag is usually an ISO 639-2 value and indicates the linguistic variety; the second one is an ISO 3166-1 region code. For example, *es-MX* stands for Spanish as spoken in Mexico, *es-419* for Spanish as spoken in Latin America. In addition, the ISO 639-3 three-letter language codes and ISO 15924 codes are used. One can specify, for instance, that the language being used in a particular encoded element is in the Cantonese dialect (gan) of Chinese (zh) as spoken in Hongkong (HK) and written in Latin characters (Latn): these subtags have to be arranged in the proper order: *zh-gan-Latn-HK*.

49 While identifiers for standard linguistic varieties are adequate for many text encoding projects, some of our projects in variational linguistics, especially dialectology, need to provide locational granularity beyond what is specified in the second subtag. To solve this problem, ICLTT staff make use of private use subtags (which, according to BCP 47, must be introduced with an *x* singleton). They help to indicate particular geographical locations and writing systems that cannot be identified by one of the standards referenced by BCP 47. Consider the following case of the representation of the lemma for Egyptian Arabic *book*:

```
...
   <form type="lemma">
      <orth xml:lang="ar-arz-x-cairo-vicav">kit#b</orth>
   </form>
...
```

50 In constructing these labels, ISO standards have been applied wherever possible. The value of the BCP 47 language tag (that is, the value of the @xml:lang attribute) starts with the shortest available ISO 639 code: *ar* stand for Arabic. This is followed by an extended language subtag. ISO 639-3 provides 30 identifiers for what in the specification is called *individual languages*, which all belong to the macrolanguage Arabic.[24] The three-letter subtag *arz* translates into Egyptian Arabic.[25] Unfortunately, this is not precise enough for purposes of dialectology, as the dialects spoken in Egypt are subdivided into a great number of quite divergent dialects, which our system has to accommodate (with private use subtags, as explained above). The schema we are using constructs these subtags from two components: location and writing system. The first component (location) does not require further explanation, whereas the second component (writing system) in this example is *vicav*, which stands for *Viennese Corpus of Arabic Varieties (transcription)*, a hybrid system for transcription that attempts to represent the most common current usage in the community. While this system of constructing language labels has served our purposes very well, for documentary purposes it is still recommended to specify the exact meaning of the toponym (the first component of our private use subtag) in the <teiHeader> of the dictionary.[26] We hope that future standards for language tags will allow for geo-spatial references with much finer granularity.

51 The following example is taken from a Modern Persian dictionary entry, the English translation of the lemma is ‘to go, to walk’.

```
...
   <form type="lemma">
      <orth xml:lang="fa-Arab">####</orth>
      <orth xml:lang="fa-x-modDMG">raftan</orth>
   </form>
...
```

52 The two letters *fa* identify the language (Modern Persian, ISO 639-2), and *Arab* indicates the writing system (ISO 15924).[27] The private use subtag indicates the system used to transcribe the Arabic characters. In this particular case, *modDMG* is a ***mod****ified version of the system of the* ***D****eutsche* ***M****orgenländische* ***G****esellschaft*. Documentation of the system and the applied modifications are explained in the dictionary’s <teiHeader>.

## 5.5. Etymologies

53 The encoding of etymologies is straightforward in TEI. As in canonical TEI, our schema allows the <etym> element as a child of entry. <etym> in turn contains one or more <lang> elements. To make the information inside the <lang> element explicit, a @sameAs attribute is added whose value points to feature structures referring to an ISO 639-2 value.

```
...
<etym><lang sameAs="#iso2_la">Latin</lang></etym>
...
```

## 5.6. Adding Semantics

54 So far, we have discussed phenomena pertaining to orthography and morphology, but we have not yet touched on equivalents or translations of the lemmas. All of this kind of information

is placed in one or more <sense> elements. In monolingual dictionaries, equivalents of the lemma are encoded as <def> elements. Definition in this particular sense implies synonym or paraphrase. When working on bi- and multi-lingual data, translations are encoded as <cit> elements, and the content proper is placed in <quote> elements within these.[28] Translations in more than one language are encoded by means of several <cit> elements.

```
<entry>
   <form type="lemma"><orth>Schloss</orth></form>
   <sense>
      <cit type="translation" xml:lang="en">
         <quote>castle, palace</quote></cit>
      <cit type="translation" xml:lang="fr">
         <quote>château, palais</quote></cit>
   </sense>
...
```

55 In addition to the <def> and <cit> elements, our schema only allows <gramGrp> and <usg> inside the <sense> element.

```
...
   <sense>
      <usg type="dom">colour</usg>
      <cit type="translation" xml:lang="en">
         <quote>black</quote></cit>
   </sense>
...
```

### *5.6.1. Grammatical Valency*

56 The appropriate encoding of grammatical phenomena often called *valency* or *government* is still not entirely resolved in the TEI Guidelines. The Guidelines provide only two examples for the <colloc> element; both are encoded with a @type attribute that has the value *prep* (for preposition). One is an entry for French *médire de*, which in English translates as "to speak ill of".

```
<entry>
   <form>
      <orth>médire</orth>
   </form>
   <gramGrp>
      <pos>v</pos>
      <subc>t ind</subc>
      <colloc type="prep">de</colloc>
   </gramGrp>
</entry>
```

57 The second example is an entry with Chinese *shuō* "to speak" as lemma, followed here by the resultative particle *dào*, which can be rendered in this context as *of* or *about*.

```
<entry>
   <form>
      <orth>#</orth>
   </form>
   <gramGrp>
      <colloc type="prep">#</colloc>
   </gramGrp>
</entry>
```

58 The solution we had in mind was something that would reach beyond what, to a majority of linguists, would be acceptable as *collocate*. For this reason, we decided to consider other encoding options.

59 A uniform system for specifying a lexical item's main complements (arguments in linguistic nomenclature) was needed. Note that this part of our encoding system is still in its infancy. However, it is important to mention that this kind of information is invariably marked up within the <sense> element. Our current encoding is illustrated by the following excerpt:

```
...
```

```
   <sense>
      <gramGrp>
         <gram type="argument">in</gram>
      </gramGrp>
      <cit type="translation" xml:lang="en">
         <quote>sich interessieren (für)</quote></cit>
   </sense>
...
```

60 In our customization, the <gram> element is used to list selected arguments relevant to the material of a specific project. None of the projects aims at the exhaustive coverage of arguments. We have also been thinking about making use of feature structures, as in the following example:

```
...
   <fs type="syntacticBehaviour">
      <f name="coreArguments" feats="#optSubj #oblPrepObj "/>
   </fs>
...
```

61 The above structure will appear very familiar to readers conversant with LMF (Lexical Markup Framework). With a generic solution designed along these lines, a precise expression of *valency* or *government* is achievable. It would also be feasible to differentiate between mandatory and optional arguments.

### *5.6.2. Dictionary Examples*

62 As explained above, all ICLTT dictionary projects are tightly interlinked with corpus-building activities. For this reason, the encoding of examples in dictionary entries requires particular attention. The relation between dictionary and corpus has to be seen as bidirectional: on the one hand, lexicographic data are designed to be used in the analysis of corpora, yet on the other hand, corpora are used to enhance and refine dictionaries.

63 One important requirement was identified at the outset of our work: dictionary examples must be reusable in different entries of a dictionary. As we did not want to duplicate data in the dictionary, the natural choice was to work with <ptr> elements to reference examples.

64 In TEI P5, dictionary examples are encoded as <cit> elements with @type attributes. Except for the value of the @type attribute, they look exactly like translations. The following example is taken from an isiZulu-English glossary:

```
...
   <cit type="exampleSentence" xml:>
      <quote>Amanzi ayabanda.</quote>
      <cit type="translation" xml:lang="en">
         <quote>The water is cold.</quote>
      </cit>
   </cit>
...
```

65 In our TEI-encoded dictionaries, examples such as the one above are children of the <body> element. Our dictionary editing program organizes dictionaries into three basic units—one metadata record (a <teiHeader> element) for the whole dictionary, an open number of entries, and dictionary examples (which can either be multi-word expressions, phrases or sentences with respective translations)—each of which are stored as separate database entries. Examples can then be linked to particular <sense> elements through a unique identifier which is referenced via the @target attribute of a <ptr> element:

```
<entry xml:>
   <form type="lemma">
      <orth>amanzi</orth>
   </form>
      ...
   <sense>
      <cit type="translation" xml:lang="en">
         <quote>water</quote>
      </cit>
```

```
      <ptr type="exampleSentence" target="#amanzi_ayabanda_01"/>
   </sense>
</entry>
```

66 Usually, one example <cit> element contains a single <quote> element. Nevertheless, in some cases multiple <quote> elements might be required, such as to give the example in several orthographic representations (with the @xml:lang attribute differentiating them). The following example is again taken from the Colloquial Cairene Arabic dictionary:

```
...
   <cit type="exampleSentence" xml:>
      <quote xml:lang="ar-arz-x-cairo-vicav">id-dinya #arr# #awi in-nahar-da.</quote>
      <quote xml:lang="ar-arz-x-cairo-modDMG">id-dinya# #arr ’awi in-nahar-da.</quote>
      <quote xml:lang="ar-arz-x-cairo-IPA">id-dinya #arr# ’awi in-nahar-da.</quote>
      <quote xml:lang="ar-arz-Arab-x-cairo">.######## ### ## ######</quote>
      <cit type="translation" xml:lang="en">
         <quote>It’s very hot today.</quote>
      </cit>
   </cit>
...
```

## 5.7. Metadata at the Level of the Dictionary Entry

67 Recording production metadata has been a recurring issue in many of the ICLTT’s encoding projects, and the lexicographic work is no exception. It is common knowledge that the TEI provides very efficient mechanisms to make statements about all kinds of responsibility in the <teiHeader> element. However, problems arise when such statements are needed on a more granular level than the whole TEI document.[29] In parts of our lexicographic work, we need to make responsibility statements not only about the whole dictionary but also about particular entries.

68 In everyday lexicographic work, it is not enough to assign the ID of one single lexicographer to an entry; one might want to trace who did what and at what time. As neither <revisionDesc> nor <change> may be used as child elements of <entry>, we considered various options to accommodate this information in our TEI structures. The intention was not to store production-related metadata only as a separate field in the database but to preserve this data in a self-contained manner as part of the entries so that this data would be passed on whenever a digital dictionary gets distributed.

69 Two elements were singled out which appeared to be plausible candidates to handle metadata about revisions of entries:  and <note>. These elements both have sufficiently generic semantics and, most importantly, may be used as children of the <entry> element. We first tried to encode metadata on revisions like this:

```
...
   <note type="revisionDesc">
      <list>
         <item><date when="2011-10-11"/>charly, added POS</item>
      </list>
   </note>
...
```

70 We wanted to stay as close as possible to comparable TEI structures without bending the semantics of particular elements. We decided in favor of a  element for revisions, containing a *feature structure*. This  element is inserted as the last element at the end of the entry. Each modification of the entry is registered by means of an <fs> element:

```
...
   <div type="revisionDesc">
      <fs type="change">
         <f name="who">charly</f>
         <f name="when">2011-10-15</f>
         <f name="what">added POS</f>
      </fs>
      ...
</div>
...
```

71 The <fs> element corresponds to the TEI <change> element, and the single features (<f> elements) correspond to the attributes of @change. Such constructs can also be used to register status information: labels carrying values such as *proposal*, *draft*, and *approved* can be used to control release of selected entries to the public.

## 6. Tools

72 So far, work on these digital lexical resources has been accomplished using a software application developed in-house. The program was initially used in collaborative glossary editing projects carried out as part of language courses at the University of Vienna. As it proved to be flexible and adaptable enough, it has been put to use in the ICLTT's dictionary projects.

73 At the heart of the software application is the dictionary editing client, a standalone application temporarily dubbed the *Viennese Lexicographic Editor* (VLE). It supports web-based editing and dictionary entries are stored on a web server. All additional software components (PHP and MySQL) are open-source and freely available. Communication between the dictionary client (VLE) and the server has been implemented as a RESTful web service.

74 While the dictionary editor is geared towards general use with XML data, it is particularly suitable and customized for the use with TEI-encoded data. In addition to fully customizable XSLT stylesheets, the tool includes a number of helpful built-in features described in brief below.

75 Configurable keyboard layouts are designed to support the input of Unicode characters usually not available in standard key assignments. Recent VLE versions allow the automatic assignment of a keyboard to particular @xml:lang attributes to spare users of manual switching between keyboard layouts. For example, when the user works on contents of an element provided with an @xml:lang="ru" attribute, VLE automatically activates the Russian keyboard layout; on entering an element with the attribute @xml:lang="de", it switches back to the German layout.

76 Entry-specific metadata can be generated automatically whenever an entry is saved. IDs of both entries and examples are created automatically on the basis of the contents of the respective items.

77 Another feature of the dictionary editor is a special module that assists with the integration of corpus examples into dictionaries. The principal idea behind this module was optimizing access to digital corpora: the corpus interface of the dictionary writing application enables lexicographers to launch corpus queries and insert them into existing dictionary entries without using the clipboard to copy-and-paste, which would inevitably result in a lot of inefficient typing or clicking.[30]

78 The validation of our dictionary data currently uses XML Schema, but the most recent versions of VLE have been delivered with a newly integrated library that is also capable of validating the data against RelaxNG schemas.

## 7. Conclusion

79 The heterogeneity of linguistic annotation has been and will remain a major obstacle for interoperability and reusability of language resources. Over the past few years, there has been increased awareness among developers and users of the need to achieve a higher degree of convergence in many parts of their encoding systems. ICLTT staff members' previous experiences with LMF have shaped the TEI customization, and the draft MAF specification is significantly influencing linguistically motivated TEI applications. In creating digital dictionaries, both of these ISO specifications (and others referenced by them) will continue to complement the work with the TEI Guidelines.

80 All of our lexicographic endeavors have been guided by a vision of an ever more densely knit web of dictionaries and more reusable, standards-based, and ideally publicly available language resources. Such resources and the respective tools for creation and access form an integral part of state-of-the-art ICT infrastructures. The ICLTT's interest in furthering the outreach of the TEI and integrating the Guidelines into the newly evolving digital infrastructures has, among others reasons, been motivated by their strong commitment to the European infrastructure projects CLARIN and DARIAH.

81 In conclusion, we would like to emphasize that our customization of the TEI P5 dictionary module has proved to be a solid foundation for new lexicographic projects. While there is no doubt that much work remains to be done, we strongly believe that the results of our experiments furnish ample evidence that TEI P5 can not only be used to represent digitized print dictionaries but also for NLP purposes.

---

### *Bibliography*


Atkins, Beryl T.S., and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford; New York: Oxford University Press.

Banski, Piotr, and Beata Wójtowicz. 2009. "FreeDict: An Open Source Repository of TEI-encoded Bilingual Dictionaries". Paper presented at the 2009 Conference and Members' Meeting of the TEI Consortium, Ann Arbor, Michigan, November 9–15, 2009. http://www.tei-c.org/Vault/MembersMeetings/2009/files/Banski+Wojtowicz-TEIMM-presentation.pdf.

Bel, Nuria, Nicoletta Calzolari, and Monica Monachini, eds. 1995. "Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets". MULTEXT Deliverable D1.6.1B. Pisa.

Budin, Gerhard, Heinrich Kabas, and Karlheinz Moerth. 2012. "Towards Finer Granularity in Metadata: Analysing the Contents of Digitised Periodicals". In *Journal of the Text Encoding Initiative* 2. doi: 10.4000/jtei.416.

Budin, Gerhard, and Karlheinz Mörth. 2011. "Hooking up to the Corpus: the Viennese Lexicographic Editor's Corpus Interface". In *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10–12 November 2011*, edited by Iztok Kosem and Karmen Kosem. Ljubljana: Trojina, 52–59. Institute for Applied Slovene Studies.

Dobrovolsky, Dmitry O. 2008–2010. *Neues Deutsch-Russisches Grosswörterbuch*. 3 vols. Moscow: AST.

Faith, R. 1997. *A Dictionary Server Protocol*. http://www.rfc-editor.org/rfc/rfc2229.txt.

Farrar, Scott, and D. Terence Langendoen. 2003. "A Linguistic Ontology for the Semantic Web". *GLOT International* 7 (3): 97–100.

Hass, Ulrike, ed. 2005. *Grundfragen der elektronischen Lexikographie: Elexiko, das Online-Informationssystem zum deutschen Wortschatz*. Berlin; New York: W. de Gruyter.

Hausmann, Franz Joseph, Oskar Reichman, Herbert Ernst Wiegand, and Ladislav Zgusta, eds. 1989–1991. *Dictionaries. An International Encyclopedia of Lexicography.* 3 vols. Berlin; New York: W. de Gruyter.

Ide, Nancy, Adam Kilgarriff, and Laurant Romary. 2000. "A Formal Model of Dictionary Structure and Content". In *Proceedings of the Ninth EURALEX International Congress: EURALEX 2000*: Stuttgart, Germany, August 8th–12th, 2000, 113–126. Stuttgart: Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Ide, Nancy, Jean Veronis, Susan Warwick-Amstrong, and Nicoletta Calzolari. 1992. "Principles for Encoding Machine Readable Dictionaries". In *EURALEX '92 Proceedings: Papers Submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland*. Tampere, Finland: Tampereen Yliopisto.

ISO-24611 (Draft). 2008. *Language resource management — Morpho-syntactic annotation framework*.

ISO-24613. 2008. *Language resource management – Lexical markup framework (LMF)*.

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2009. "ISOcat: Remodelling Metadata for Language Resources". In *International Journal on Metadata, Semantics and Ontologies* 4: 261–276.

Mörth, Karlheinz. 2002. "The Representation of Literary Texts by Means of XML: Some Experiences of Doing Markup in Historical Magazines." In *Digital Evidence. Selected Papers from DRH 2000, Digital Resources for the Humanities Conference*, edited by Michael Fraser, Nigel Williamson, and Marilyn Deegan, 17–32. London: Office for Humanities Communication.

Romary, Laurent, Susanne Salmon-Alt, and Gil Francopoulo. 2004. "Standards Going Concrete: From LMF to Morphalou". In *Workshop on Enhancing and Using Electronic Dictionaries*. Geneva: Coling.

Romary, Laurent. 2010. “Standardization of the Formal Representation of Lexical Information for NLP”. In *Dictionaries: An International Encyclopedia of Lexicography*. Supplementary Volume: *Recent Developments with Special Focus on Computational Lexicography*. http://arxiv.org/abs/0911.5116.

Romary, Laurent. 2010. “Using the TEI Framework as a Possible Serialization for LMF”. Paper presented at *RELISH workshop, August 4–5, 2010, Nijmegen, Netherlands*. http://hal.archives-ouvertes.fr/docs/00/51/17/69/PDF/NijmegenLexicaAugust2010.pdf.

Sarup, Lakshman. 1920–27. *The Nighantu and the Nirukta: The Oldest Indian Treatise on Etymology, Philology and Semantics*. Delhi.

Snell-Hornby, Mary. 1986. “The Bilingual Dictionary: Victim of its own Tradition?” In *The History of Lexicography*, edited by Reinhard Hartmann, 207–218. Amsterdam: John Benjamins.

TEI Consortium. 2012. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.1.0. Last updated June 17. N.p.: TEI Consortium. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Wegstein, Werner, Mirjam Blümm, Dietmar Seipel, and Christian Schneiker. 2009. “Digitalisierung von Primärquellen für die TextGrid-Umgebung: Modellfall Campe-Wörterbuch“. http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R4_1.pdf.

Wilkinson, Endymion. 2000. *Chinese History. A Manual*. Cambridge, Mass.: Harvard University Asia Center.

---

### *Notes*

1 There is no reliable information available as to his date of birth. Tradition assumes the 5th or 6th century BC. See Sarup (1920–27, 54).

2 While none of these works can be regarded as an absolute first, they can all be seen as important milestones in their respective traditions.

3 A project working on Russian Wiktionary versions is the Wiktionary-Export project which also produces TEI versions (http://wiktionary-export.nataraj.su/en/about.html).

4 http://www.ilc.cnr.it/EAGLES96/home.html

5 http://www.olif.net/

6 http://www.w3.org/2001/sw/BestPractices/WNET/ISLE_D2.2-D3.2.pdf

7 http://code.google.com/p/lift-standard/

8 http://www.w3.org/TR/owl-ref/

9 http://tools.ietf.org/html/rfc2229

10 A standardized object-oriented modeling language.

11 The ICLTT’s dictionary editor VLE provides a tool to convert some of the TEI encoded dictionary data into LMF. This end is achieved by making use of XSLT stylesheets to transform the TEI data into an XML format that looks very much like the XML serialization as it can be found in the ISO specification.

12 This also shows in the fact that the P4 chapter was titled “Print dictionaries”, whereas the current P5 version bears the title “Dictionaries”.

13 An example of what we would like to see more of can be found on the ICLTT’s experimental Showcase website: http://corpus3.aac.ac.at/showcase/index.php/dictionary. In this dictionary interface, each entry can also be viewed with its TEI encoding.

14 Among the well-documented examples of TEI P5 encoded dictionaries, there is the CAMPE dictionary, a product of the TextGrid project (Wegstein 2009). While most data in the field are not easily available, let alone for reusing or further development, a number of P5-compliant dictionaries were made freely available by the FreeDict project (Banski 2009).

15 See the LMF website: http://www.lexicalmarkupframework.org/.

16 The general structure of these items of lexicographic information has been discussed in various publications before. See Ide et al. (1992), Ide et al. (2000), and Romary (2011).

17 These are <case>, <colloc>, <def>, <etym>, <form>, <gen>, <gramGrp>, <hom>, <hyph>, <iType>, <lang>, <lbl>, <mood>, <number>, <oRef>, <oVar>, <orth>, <pRef>, <pVar>, <per>, <pos>, <pron>, <re>, <sense>, <subc>, <superEntry>, <syll>, <tns>, <usg>, and <xr>.

18 ISO-24613:2008(E), 39.

19 Feature structures are a general-purpose data structure that have become a widely used means of representation in linguistics. They have a longstanding tradition in the TEI. A chapter on the topic in the TEI Guidelines goes back to P3 (Sperberg-McQueen and Burnard 1994, 394–431).
20 http://wiki.verbix.com/Documents/VerbixLanguageCodes
21 Global attributes can be used on all elements of the TEI encoding scheme.
22 BCPs are published by the Internet Engineering Task Force together with RFC (request for comments) documents.
23 BCP 47 is made up of two IETF documents: RFC 4646 and RFC 4647. A good overview is given in TEI Consortium 2012, liv.
24 The registration authority for ISO 639-3 is SIL International (http://www.sil.org/iso639-3/codes.asp).
25 It is interesting that W3C discourages the use of macrolanguage subtags (http://www.w3.org/International/questions/qa-choosing-language-tags.en#langsubtag). The label *arz-x-cairo-vicav* would be as clear as *ar-arz-x-cairo-vicav*.
26 While Cairo, Illinois (USA), will probably not be confused with the Egyptian capital in this context, other ambiguities will definitely occur.
27 The language identifier *fa* has the "Suppress-Script: Arab" entry set in the IANA registry. That means that it is the default and should be omitted. However, we decided to be more explicit in such cases as the different <orth> elements are being used in our markup scheme exactly for the purpose of representing different writing systems.
28 The structure of the <sense> block has been heavily affected by the transition from P4 to P5. The <trans> and <tr> elements have been removed from P5.
29 In a paper presented at the TEI Members' Meeting last year, we discussed the possibility of assigning TEI headers through links to particular divisions of text documents (Budin and Moerth 2011).
30 See Budin 2011.

***Cite this article***

Electronic reference

Gerhard Budin, Stefan Majewski and Karlheinz Mörth, « Creating Lexical Resources in TEI P5 », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 05 November 2012. URL : http://jtei.revues.org/522 ; DOI : 10.4000/jtei.522

***Authors***

**Gerhard Budin**
Gerhard Budin is full professor for terminology studies and translation technologies at the Centre of Translation Studies at the University of Vienna, director of the Institute for Corpus Linguistics and Text Technology of the Austrian Academy of Sciences, member (kM) of the Austrian Academy of Sciences, and holder of the UNESCO Chair for Multilingual, Transcultural Communication in the Digital Age. He also serves as vice-president of the International Institute for Terminology Research and Chair of a technical sub-committee in the International Standards Organization (ISO) focusing on terminology and language resources (ISO/TC 37/SC 2 2001–2009, SC 1 2009-present). His main research interests are language technologies, corpus linguistics, and knowledge engineering, E-Learning technologies and collaborative work systems, distributed digital research environments, terminology studies, ontology engineering, cognitive systems, cross-cultural knowledge communication and knowledge organization, philosophy of science, and information science.
**Stefan Majewski**
Stefan Majewski studied English Language and Literature as well as Sociology at the University of Vienna and Electronics at the Vienna University of Technology. He graduated in English Linguistics with a focus on research infrastructures for corpus linguistics. Currently, he is working at the Austrian Academy of Sciences, where he coordinates and works for the "Data Service Infrastructure for the Social Sciences and Humanities" (DASISH) project. He is also employed by the Göttingen State and University Library, where he works for the "TextGrid" project in research and development. His current interests focus on research infrastructures and annotation systems.

**Karlheinz Mörth**
Karlheinz Mörth is senior researcher and project leader at the Institute for Corpus Linguistics and Text Technology (ICLTT) of the Austrian Academy of Sciences, lecturer at the University of Vienna and co-head of the DARIAH Virtual Competency Centre 1 (eInfrastructure). Proceeding from a broad background in cultural, literary and linguistic studies, he has been working on a number of scholarly digital projects. He has contributed to the design and creation of the Austrian Academy Corpus (AAC), taking responsibility for text encoding and software development. His current research activities focus on eLexicography and text technology for linguistic research.



***Abstract***

Although most of the relevant dictionary productions of the recent past have relied on digital data and methods, there is little consensus on formats and standards. The Institute for Corpus Linguistics and Text Technology (ICLTT) of the Austrian Academy of Sciences has been conducting a number of varied lexicographic projects, both digitising print dictionaries and working on the creation of genuinely digital lexicographic data. This data was designed to serve varying purposes: machine-readability was only one. A second goal was interoperability with digital NLP tools. To achieve this end, a uniform encoding system applicable across all the projects was developed. The paper describes the constraints imposed on the content models of the various elements of the TEI dictionary module and provides arguments in favour of TEI P5 as an encoding system not only being used to represent digitised print dictionaries but also for NLP purposes.

***Index terms***

***Keywords :*** P5, dictionaries, digital lexicography, NLP

***Author's notes***

This paper is based on a presentation given at the TEI Members' Meeting 2011 in Würzburg, Germany.