Slovenski jezik – Slovene Linguistic Studies 9 (2013): 163–183

Nina Ledinek and **Andrej Perdih**
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Ljubljana

Creating XML Schemas for Lexicographical Projects: the Case of the Dictionary of the Slovene Literary Language of the 16th Century

Članek obravnava strategije oblikovanja XML-shem za slovarske projekte na primeru kompleksne strukture Slovarja slovenskega knjižnega jezika 16. stoletja in pojasnjuje, kako različne možnosti izdelave XML-sheme vplivajo na leksikografsko delo in kasnejšo uporabo slovarja. Prikazani so različni splošni vidiki izdelave shem za slovarske projekte, ki jih je smiselno upoštevati: vsebinski, praktični in tehnični vidik. Na tej osnovi je utemeljena odločitev za način formalnega opisa strukture v danem računalniškem formatu.

The article discusses the strategies applied for creating XML Schemas in dictionary projects, based on experience gained during creation of the schema of the Dictionary of the Slovene Literary Language of the 16th Century, and it explains how the different possibilities of the construction of XML Schemas influence the lexicographical work and the subsequent use of the dictionary. Three general aspects to be considered in designing schemas for dictionary projects are described in the article: the content aspect, the practical aspect, and the technical aspect. On this basis, the decision on the formal description method of the structure in a given schema definition language is made and justified.

1 Introduction

In a dictionary database written in the currently widely adopted XML format, the structure of the lexicographical data is described by a schema. Regardless of the schema definition language (DTD, XML Schema, RELAX NG) supported by the lexicographical software, it is essential to create a schema that, firstly, corresponds to the lexicographical content requirements, secondly, adequately structures the data set, and, thirdly, enables the employment of the program's solutions, which facilitate and accelerate the work done by lexicographers. If the construction of the schema leans towards maximizing the simplicity, transparency, and intuitive quality of the structure, while always showing consistency with the conceptual requirements, lexicographers can more easily focus on its contents and lose less time with technical issues. Moreover, dictionary users will subsequently find advanced searches to be more straightforward and easier to use. It is also important to take into consideration the subsequent use of data in other reference works as well as for language technology needs or data conversion. The different schema implementations of the chosen dictionary structure can impact any one of these factors. For these reasons it is especially important to devote a considerable amount of time to the construction of the schema and the refinement of the concept in the case of dictionaries containing large amount of different types of data. Last but not least, the creation of schemas usually reveals gaps in the adopted dictionary concept as regards formal requirements, and may also disclose content uncertainties of the designed dictionary concept.

The COBUILD project was one of the first lexicographical projects, dating back to the early eighties of the last century, in which the data was tagged in a similar way as today, and in which the lexicographical material was prepared with the help of specialized electronic lexicographical tools. The lexicographical database was structured and written in a standard textual format ASCII (cf. Clear 1987: 51, Krek 2004: 4). The fact that standards for encoding lexicographical information have lately been established shows that modern lexicography was significantly influenced particularly by the fact that lexicographers usually do not perceive a dictionary (initially) as a book anymore, but rather as an extendable machine-readable database structure, which can be used for different purposes and where the data is organized in a hierarchical order, marked (according to the standard) and interconnected, thus enabling the interchangeability of data, different content views, and a much faster and more accurate lexicographical data search (cf. Abel 2012, Smrž 2001).

In the core part of this article we look at three aspects that should be considered when creating a schema: the content aspect, the practical aspect, and the technical aspect. In section 4 we demonstrate two of the possible solutions for designing XML Schemas of the morphological header in the Dictionary of the Slovene Literary Language of the 16th Century while considering these three aspects. Furthermore, we consider and evaluate the two solutions based on the advantages and disadvantages of their use, and, finally, we justify our selection of the most suitable structure.

2 Using XML Format in Lexicography

Most dictionaries are known for their relatively complex structure (Abel 2012: 84), due to the fact that they include a large amount of various linguistic data. Different functions of the individual segments of dictionary entries are usually pointed out via different text style, the use of symbols and punctuation, as well as clearly defined subdivisions of dictionary entries. Even dictionaries with relatively simple microstructure usually contain several types of data, e.g., lemma, part of speech, and irregular forms, whereas dictionaries with relatively complex microstructure usually include additional information such as inflections, pronunciation, lexeme usage, collocations, synonyms, phraseology, etymology, frequency of usage, etc. Due to both an enormous amount of diverse data and the need for its transparency, it is essential that the different language data types in hierarchically structured machine-readable dictionary databases are clearly marked with unique and meaningful tags.

The standard mark-up language suitable for marking dictionaries and other language data, which has established itself in the past years is XML (eXtensible Markup Language). The tree structure of the XML file is suitable for encoding hierarchically structured data, thus allowing the user to apply advanced searches of the database as well as a wide variety of complex types of data processing. Unlike HTML, metatags in XML format are not specified in advance, but can be set by the user almost arbitrarily. By default, Unicode character encoding is used, which, together with the appropriate Unicode font, provides ample opportunity for the use of a vast range of characters. XML files are used by a number of lexicographical programs. At the Fran Ramovš Institute of Slovene Language SRC SASA iLEX is being used.[1] Among the well-known ones are also ABBYY Lingvo Content,[2] IDM DPS[3] and TshwaneLex.[4] Termania[5] has been developed in Slovenia. XML files are plain text files. Their greatest virtue is that they are transferrable between different programs, operating systems, and devices, which is essential in the long term, because the dictionary written in XML format can be used in any program that is able to read plain text files.

XML format is especially suitable for structuring data types and their long term storage by way of allowing the data to be portable and interchangeable. However, it is not primarily intended for the display of data. For a comprehensive use of XML files, various XML languages were developed inside the XML family. XSLT language is meant for converting XML documents into other formats, which can include changing the XML document or converting it into formats suited for screen display or for print. HTML and XHTML are both formats used for screen display (the most common example is web sites), whereas XSL:FO is suitable for converting data into PDF, i.e., print. These kinds of conversions are almost indispensable for the user, because, as mentioned before, XML is a markup language and not easy to read, whereas people want to see a formatted text on screen or on paper. XQuery is used for data search in XML files; XPath is used for navigation, while the structure of XML files is defined by different schema languages: DTD (.dtd), XML SCHEMA (.xsd), and RELAX NG (.rng). Schematron represents a different type of schema, which verifies if the content of an element is allowed, depending on the content of another element (cf. Berglund 2006, Birbeck et al. 2010, Bray et al. 2006, Hunter 2007).

[1] http://www.emp.dk/

[2] http://www.abbyy.com/lingvo_content/

[3] http://www.idm.fr/products/dictionary_writing_system_dps/27/

[4] http://tshwanedje.com/tshwanelex/

[5] http://www.termania.net/

3 The Dictionary's Structure and the Schema for XML Files[6]

The creation of a dictionary concept or the appropriate structure of dictionary entries is one of the most challenging steps before commencing with editorial work. It is crucial to design a dictionary structure, which defines elements of the dictionary entries and the allowable sequences in which they can occur and as such serve as a basis for any type of dictionary entries, despite the obvious fact that not all of their available structure components will appear in every dictionary entry. The task of creating such a structure is usually especially demanding for certain types of specialized dictionaries (as is the case with the dictionary presented in this article), in particular for dictionaries that aspire to display a variety of different data.

The schema describes the formal structure of the dictionary database in XML format. It determines which elements are allowed in the dictionary database, the hierarchical relations between them and their order, the possibilities of their different combinations or exclusions, as well as the number of element occurrences when we wish to specify more than one consecutive identical element (Hunter et al. 2007). The schema also provides the formal content requirements for elements: whether any kind of content is allowed, or, if there is a restriction to the list of possible choices (drop-down menu), or, if there is a length restriction regarding the content, if the content is restricted to digits, if elements can have attributes, etc. A schema is therefore a kind of transformation of a dictionary concept into computer language. As regards the dictionary content, though, it helps only with certain technical requirements and can in no case prevent content inadequacy or inconsistency with the conceptual guidelines or the linguistic reality.

There are several standard schema definition languages to describe XML documents; the two most established and widespread today are DTD and XML Schema.[7] Even though DTD is slightly less popular than XML Schema, which enables a somewhat more flexible and detailed description, using one or the other schema definition language in practice usually depends on the lexicographical program, taking into account that some programs do not support different schema definition languages.[8]

For various reasons, such as interchange, longevity, simpler processing, global standards have been established in regard to language data format for encoding

[6] The lexicographical material described below is the analyzed material being edited by the research group composed of Kozma Ahačič, Metod Čepar, Alenka Jelovšek, Andreja Legan Ravnikar, Majda Merše, Jožica Narat and France Novak. The XML Schema was prepared in close cooperation with Kozma Ahačič.

[7] *XML Schema* (.xsd file) and *schema* are not synonymous, because *XML Schema* is one of the schema definition languages just like *DTD*, which describes the structure of the XML file content (Hunter et al. 2007: 145). For more information on XML Schema see Thompson et al. 2004.

[8] IDM DPS and Tshwanelex support DTD, iLEX supports RELAX NG and XML Schema, Termania supports XML Schema.

dictionary or lexicon databases. Here we will mention Lexical Markup Framework (LMF),[9] Lexical Interchange Format Standard (LIFT),[10] TEI Guidelines.[11] This list, however, is not exhaustive. Despite the fact that the data could be written in at least one of the standard formats, we have not yet decided to take this step. The decision not to follow any of these standards was based on the judgment that it would be less sensible to consider this kind of format, due to the extreme complexity of the dictionary's microstructure, which would cause great difficulty in following a standard format. Despite its flexibility it would be difficult to follow the standard encoding. A question to take into consideration is also whether the dictionary encoding standards will be altered by the time the dictionary is finished. In making our decision we considered the fact that the dictionary database is primarily intended for specialized users for language research, whereas to a lesser extent it will also be useful for 16th century Slovene texts processing. Also, the dictionary concept was designed in a time when these standards were not yet ubiquitous or established and therefore adjusting the structure to any of the standards is more time consuming.

Besides the fact that the software for lexicographical work enables editors to edit the language data and visually represent it, it is one of the basic tasks of the software to ensure consistency between the dictionary entries and the schema and to point out the irregularities in the structure as well as the formal content of the elements. It is therefore a lexicographical tool, which helps to ensure the consistency of individual dictionary entries and the entire structure of the dictionary. This is especially crucial for larger projects, in which numerous lexicographers are involved. Even though the (base) schema is created before the editing process takes place, it is customary that the schema is partially modified and elaborated during the preparation of the first (test) dictionary entries. Lexicographers and computer experts must be attentive to any possible inadequacies in the dictionary structure and consequently the schema, for in doing so they can eliminate any confusion, point out the technical shortcomings, or coordinate any different existing views on the content of the emerging dictionary.

In the process of preparing XML Schemas for dictionaries at the Fran Ramovš Institute for Slovene Language SRC SASA it has become clear that it is reasonable to consider several aspects, which can affect the creation of XML Schemas. These aspects are:

- the content aspect,
- the practical aspect,
- the technical aspect.

[9] http://www.lexicalmarkupframework.org/

[10] http://code.google.com/p/lift-standard/

[11] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

3.1 The Content Aspect

The content aspect is in fact a conceptual requirement for the dictionary structure, meaning that it should be projected as perceived by lexicographers and as directly as possible onto an intuitively understandable hierarchical data structure, which is defined by a schema. In practice, the aspect is most clearly articulated when determining and designating the dictionary parts or the schema elements and their allowed content and when defining the hierarchical relations between elements. The content aspect is the critical aspect to be considered when it comes to the integration of multiple elements into the parent element. This is true if it corresponds to the lexicographical view of the structure, if it is useful for structuring the data, or for transparency purposes. To put it simply, the content aspect is most evident when making the decision whether, for instance, the element header should contain all the sub elements the lexicographer sees in the header but not the ones he sees in the section where word senses are described.

An example of a simple XML document is the dictionary entry *almužna* shown below, whose content is taken from the analyzed material of the Dictionary of the Slovene Literary Language of the 16th Century:[12]

```
<hom>
   <kazalčno_geslo>
      <iztočnica>almužna</iztočnica>
      <glej_geslo>almožen</glej_geslo>
      <kazalčno_pojasnilo>rod. ed.</kazalčno_pojasnilo>
   </kazalčno_geslo>
</hom>
```

Figure 1

```
<hom>
   <reference_entry>
      <lemma>almužna</lemma>
      <cf_entry>almožen</cf_entry>
      <reference_note>Gen. sg.</reference_note>
   </reference_entry>
</hom>
```

Figure 1: English translation

Five different elements are used in this case, where the type of their content is already clear from their description. Each element starts with a start tag, e.g., <iztočnica> (lemma), and ends with an end tag, e.g., </iztočnica>; in between the tags we have the text or other hierarchically subordinate elements.

[12] The right hand indent is used to indicate child elements.

3.2 The Practical Aspect

In a dictionary with a complex microstructure, the number of different elements in a schema can quickly reach a three digit number. Certain elements can also occur in several different places at the same time, which makes sense in terms of content but can differ just enough in the relations between the elements and their order to overburden the lexicographer's memory. It is therefore reasonable to consider the practical aspect of the lexicographical schema since it must not only be logical in content but also manageable, and it must coincide with the lexicographical process as naturally as possible. The question is whether lexicographers can sufficiently manage the complex structure without it being an obstacle in their work. It is equally important, of course, that the data is arranged in a way so that the end users of the dictionary will be able to understand it.

The hierarchically deeper structure has an advantage over the flat or less hierarchically developed structure because more data combined in the parent element can be considered, thus enabling an easier search and a more detailed subsequent data processing. It is also true in this kind of structure that the data can, to a great extent, be structured according to the lexicographer’s understanding of the dictionary structure. The drawback of an excessively branched structure is, however, that the orientation within the dictionary entry is more difficult (consequently one can extend the assembly time for entries). For the most part, the lexicographer focuses his attention on the hierarch’s accuracy instead of its content; therefore, a highly complex structure can become more demanding for actual work.

3.3 The Technical Aspect

When considering how to enhance both the usability of a (completed) dictionary and the optimal amount of work put into its elaboration, the structural relationships between the elements of the schema are not of sole importance. There are other options we want to make possible. A particularly important criterion is the references to various elements in the dictionary’s microstructure, (for example if we want to make the electronic version clickable, or if we want the same data to be displayed in more than one location, if relevant), using mixed content, etc. It is also worthwhile to give some thought to the publishing of the dictionary. If the dictionary is available in electronic form only, it is not worthwhile to consider some of the adjustments a book edition might require. From the technical aspect a factor which might impact the structure of the schema is the lexicographical program itself – according to our experience, some adjustments in schema might be welcomed, for example, because the user interface influences the decision about whether it would be better to combine repetitive elements into one parent element or not. This is relevant because at one point we want to shrink the content on the screen or hide certain hierarchical elements to have a better overview of the other content we want to focus our attention on at that particular time. Programs tackle transparency issues of the interface in different ways; therefore, it is indispensable to consider the technical aspect.

The hierarchical organization of elements is of key importance when searching for data, rearranging, converting, or using it in any other way, while the occurrence of the same element in different places can, due to various combinations of the occurrences, make the search for relevant data very complicated. At the same time the risk of producing inadequate search results increases. The search for existing content is an important aspect in lexicographical work, giving it insight into the dictionary's already verified solutions. It also helps to maintain the consistency of the dictionary database and is a useful supplement of the dictionary Style Guide. Once the dictionary is complete, the dictionary database can also be used for linguistic or metalexicographic research and as a starting point for the creation of other linguistic resources, which is why the method of data organization can be very important in this aspect as well.

In terms of further use of dictionary databases, the use of elements with mixed content is equally important. This kind of element can contain text and child elements along with their own proper content. One of the reasons for the use of mixed content is the entry of a differently formatted text, for example, superscript and subscript numbers (m^3, CO_2). In regard to content, the use of mixed elements is more important for elements in the structure where their sub elements and text are "mixed" together due to the very content this kind of element contains. A typical example is the etymological section, where the reconstructed forms, the meanings of the words, foreign language words and their meanings, tags for languages they originate from, and the "normal" explanatory text between them occur in a rather unpredictable sequence. From a technical point of view, caution is generally advised when using mixed content because it can lead to the subsequent processing of this data being significantly more difficult than otherwise.

4 Creating the Structure of the Morphological Header in XML Schema

Based on the three aspects presented, we will try to illustrate two different approaches applied in the construction of the morphological header of the Dictionary of the Slovene Literary Language of the 16th Century. The dictionary is one of the major lexicographical projects in Slovenia. The preparations for the lexicographical work (including collecting the material) began in 1973, and in 2001 a test fascicle of the dictionary (Merše et al. 2001) was published. In 2011 a publication of Vocabulary of Slovene Literary Language of the 16th Century (Ahačič et al. 2011) was released, which is a compilation of 16th Century Slovene words with basic morphological information. This vocabulary will be lexicographically treated in the emerging dictionary. The written language of the time however was not standardized, which renders visible the dialectical elements used by the authors, especially regarding the phonological and morphological characteristics. Furthermore, the issue of non-standardized orthography in the texts causes variation in the recording of some phonemes as well as in the use of capitalization.[13] The dictionary schema must be structured in a way to enable recording of every one of the different variants or at least to draw attention to them, at several structural levels.

The morphological header provides information about word forms of the lemma that appear in the texts, for example, the singular, dual, and plural forms for nouns in different cases, and different forms for verbs including number, person, tense, mood, etc. The writers of these entries also add references for each word about the author and the text where the word is documented, as well as notes about any oddities that appear in the word forms. Since the dictionary is intended mostly for specialized linguists, the only word forms that appear in the dictionary are ones from which the paradigmatic model the word belongs to can be established.[14] This is due to the large number of different existent word forms. For example, in the case of nouns, the first manifested form in the singular, dual and plural is indicated, whereas, the inflected forms for each declension are indicated only if they deviate from the expected paradigm, as presented in the front matter of the dictionary. The same is true for other parts of speech. From the lexicographical point of view it is therefore quite evident that we wish to produce a structure, which will enable the entry of hierarchically classified information about the morphological categories and the actual forms of headwords. In the case of adjectives, we have six levels of information in morphological header: (1) adjective > (2) indefinite form > (3) masculine > (4) singular > (5) dative > (6) entry of documented form. The same is true for other parts of speech that take on inflections. If our objectives from the lexicographical point of view are perfectly clear, there exist at least two options on how to create a schema, which has to accept all possible forms of paradigms for relevant parts of speech, along with the information about the documented forms. We expect more than 500 combinations for the morphological categories. At the same time, we aspire to have the simplest structure possible.

The number of expected theoretically possible word forms is undoubtedly significantly greater than the actual number of the documented words in the historical texts for individual lemmas; therefore, we can assume that in the structure of the dictionary entry only some of the spaces in the morphological header for individual lemmas will be complete. Even though it is useful to be aware of this fact, it must not

[13] Digitalization of 16th century texts has not been as successful as of texts printed after 1850 and is therefore more time-consuming and costly (Erjavec et al. 2011: 121). For these reasons, no electronic corpus of 16th century Slovene texts has been compiled yet and a lot of manual work is required for the compilation of the dictionary. The situation was more successful in the recent IMPACT project which resulted in (mainly) 18th and 19th century language resources (http://nl.ijs.si/imp/). A small sample of 16th century texts is also included.

[14] This solution differs from the method adopted by the test fascicle dictionary (Merše et al. 2001), where all manifested forms are recorded, because it would significantly slow down the two-decades-long dictionary production.

affect the process of creating the schema. We have applied two different approaches for creating a morphological header schema and consequently devised two: a general schema, and an explicit schema.

4.1 The General Schema

The first option is to create a simple and general schema, which provides an adequate number of hierarchically classified general data containers, i.e., elements. This means that we only need a small number of hierarchically classified elements of the schema because the distinctive value, i.e., relevant morphological category, is given in the form of text (selection from the dropdown list of options). The advantage of this kind of schema is its universality for all parts of speech and its extreme simplicity. The number of levels in the morphological header from top to bottom is 6. In addition there is a limited number of places for data entry – all the information for one form can be entered with a maximum of six (less in some parts, of course) shifts in level, depending on the part of speech. There are 5 different elements for data entry, while the number of elements inside the morphological header, which contain only child elements, is 6. It should be noted, however, that the information for the part of speech and its category is given in the normal header; therefore, it can be omitted here.

The drawback in this type of schema is, however, that the exact location where specific information can be found is not explicitly clear, which makes such schemas non-intuitive for new participants in the dictionary making process, as well as for subsequent dictionary searches. In addition, when making a dictionary the data of this type of structure is harder to maintain. Furthermore, the potential treatment of similar information should rely on information outside the morphological header; otherwise, it is not successful. General schema's biggest problems arise from the fact that it is necessary to know exactly at what level an indication is. For example, the grammatical category of number, since it may vary between different parts of speech. This can be very challenging to memorize for lexicographers and advanced dictionary users.

The following example shows a truncated record of the morphological header for the adjective *bel* (English *white*) in XML form, corresponding to the schema in question. The structure is, for transparency purposes, slightly shortened.[15]

[15] Element *i* has been added, which signifies that its content in the historical texts may be written with different letters, while in the dictionary only one letter is recorded as a generalization. This is done due to the fact that providing all possible variants of this type would result in an overload of spelling variants in the morphological section of the entry and consequently would draw user's attention to spelling instead of morphology. The information on the possible realizations of the letter in the element *i* will be given in the introduction or in the form of hints.

```
<oblikoslovno_zaglavje>
  <pregibno_skupina>
    <OZ_1>
      <seznam1>nedoločna oblika</seznam1>
      <OZ_2>
        <seznam2>m</seznam2>
        <OZ_3>
          <seznam3>ednina</seznam3>
          <OZ_4>
            <seznam4>imenovalnik</seznam4>
            <OZ_oblika>b | é/e<i>j</i>/e/ee/ee | l</OZ_oblika>
          </OZ_4>
          <OZ_4>
            <seznam4>rodilnik</seznam4>
            <OZ_oblika>b | é/e/e<i>j</i> | liga</OZ_oblika>
          </OZ_4>
        </OZ_3>
        <OZ_3>
          <seznam3>dvojina</seznam3>
          <OZ_4>
            <seznam4>imenovalnik</seznam4>
            <OZ_oblika>bela</OZ_oblika>
          </OZ_4>
            [...]
        </OZ_3>
        <OZ_3>
          <seznam3>množina</seznam3>
          <OZ_4>
            <seznam4>imenovalnik</seznam4>
            <OZ_oblika>b | e/é/e<i>j</i> | li</OZ_oblika>
          </OZ_4>
            [...]
        </OZ_3>
      </OZ_2>
    </OZ_1>
     <OZ_1>
       <seznam1>določna oblika</seznam1>
           [...]
     </OZ_1>
    </pregibno_skupina>
</oblikoslovno_zaglavje>
```

Figure 2

English translation of terms in Figure 2:

določna oblika	definite form
dvojina	dual
ednina	singular
imenovalnik	nominative
m	masculine
množina	plural
nedoločna oblika	indefinite form
oblikoslovno_zaglavje	morphological header
OZ + *number*	morphological header level
OZ_oblika	morphological header – form
pregibno_skupina	declinable – group
rodilnik	genitive
seznam	list

In this first approach at creating a schema a choice between two options has to be made on the first level: (1) *pregibno_skupina* (*declinable_group*) and (2) *nepregibno_skupina* (*indeclinable_group*).

- At (2) *nepregibno_skupina* (*indeclinable_group*) on the second level in the inflected part we have to choose *OZ_oblika (morphological header_form)*, which is the end element for the entry of the treated form, in order to be consistent with the location of the information.
- At (1) *pregibno_skupina* (*declinable_group*) we reach the only option on the second level, *OZ_1*. Within *OZ_1* on the third level, we have *seznam1 (list1)*, which contains a list of all possible information on this level. We can then choose *OZ_2*, *OZ_3* or *OZ_4*.
- If we choose *OZ_4* on the third level, we choose *seznam4* (*list4*) and *OZ_oblika* (*morphological_header_form*) on the fourth level, where we enter the actual form of the word.
- If we choose *OZ_3* on the third level, we then have *seznam3* (*list3*) on the fourth level, where we choose the information from the menu, and *OZ_4*, which on the fifth level contains *seznam4* (*list4*), and *OZ_oblika (morphological_header_form)*, where we enter the actual form of the word.
- If we choose *OZ_2* on the third level, we then have *seznam2 (list2)* on the fourth level, where we choose the information from the menu. We can then choose *OZ_3* or *OZ_4*, where both have the same structure as mentioned above.
- With all these choices multiple repetitions in sequential order are possible for *OZ_2*, *OZ_3* or *OZ_4*, given that the hierarchically superior data is applicable

to all subordinate ones, for example, for *OZ_4*, which contains the nominative form, followed by another *OZ_4*, which contains the genitive form.

The editorial process is therefore the following: after our decision on the declinability of the word in the declinable section, we first make a decision on level *OZ_1* and its *seznam1 (list1)*. Depending on the number of hierarchical data, we then continue towards *OZ_4* (we only use the intermediate *OZ_2* and *OZ_3*, if we have to enter this much data) where we find *seznam4* and the end element *OZ_oblika (morphological_header_form)*, where we enter the actual form of the headword. With the elements that have the same name, but different information in *seznam1 (list1)*, *seznam2*, *seznam3* and *seznam4* we can enter information regardless of the part of speech. The schema is therefore simple; the only valuable thing to know is how many steps it takes to enter the information about the form – which differs for different parts of speech.

4.2 The Explicit Schema

The second option is to create a structure, where the data is more strongly embedded in the hierarchy of the structure, which then simplifies complex searches and data processing. It also leaves no doubts for lexicographers in the data entry process. Contrarily to the first option, the disadvantage of this type of structure is its complexity because the number of different elements in the schema is greatly increased, which can consequently overburden the dictionary writer at first contact with the schema – the number of different places for data entry is extremely high, 552 in 30 elements with different names on the lowest hierarchical level, where the details of the manifested form are in fact entered. The difference between the numbers means that, for example, the element *imenovalnik (nominative)*, where the nominative form is entered, can occur in various places in the structure, depending on the part of speech, number, gender, etc., of the treated form. Other information, which does not constitute the actual form of the word, is given with the name of the element, which contains the given form or is its parent element (in the first schema this data is selected from the lists). The hierarchical depth of this type of schema is five or less, which is similar to the first option (six or less).[16]

The following example demonstrates the truncated record of the morphological header for the adjective *bel* (English *white*) in XML form, corresponding to the schema in question. The structure is, for transparency purposes, slightly shortened.[17]

[16] Some elements in the morphological categories can, for practical reasons, be named slightly differently than what is common practice in modern linguistic descriptions, given that older word forms in old language are, from our point of view, probably unexpected due to the fact they were not codified and due also to their transiency.

[17] Cf. footnote 14.

```
<oblikoslovno_zaglavje>
  <pridevniško>
    <nedoločna_obl>
      <m>
        <ednina>
          <imenovalnik>b | é/e<i>j</i>/e/ee/ee | l</imenovalnik>
          <rodilnik>b | é/e/e<i>j</i> | liga</rodilnik>
        [...]
        </ednina>
        <dvojina>
          <imenovalnik>bela</imenovalnik>
        [...]
        </dvojina>
        <množina>
          <imenovalnik>b | e/é/e<i>j</i> | li</imenovalnik>
        [...]
        </množina>
      </m>
    </nedoločna_obl>
    <določna_obl>
     [...]
    </določna_obl>
  </pridevniško>
</oblikoslovno_zaglavje>
```

Figure 3

English translation of terms in Figure 3:

določna_obl	definite form
dvojina	dual
ednina	singular
imenovalnik	nominative
m	masculine
množina	plural
nedoločna_obl	indefinite form
oblikoslovno_zaglavje	morphological header
pridevniško	adjectival
rodilnik	genitive

In this explicit approach of making the schema the decision for one of the eight options takes place on the first level: (1) *nepregibno (indeclinable)*, (2) *samostalniško (nounal)*, (3) *pridevniško (adjectival)*, (4) *števnik (numeral)*, (5) *glagol (verb)*, (6) *posamostaljeno (nominalized)*, (7) *izpridevniški_prisl (deadjectival_adverb)*, (8) *pregibni_povedkovnik (declinable_predicative)*.

- At (1) *nepregibno* (*indeclinable*) this is already the end element, where we enter the given form.
- At (2) *samostalniško* (*nounal*) we can choose *nesklonljivo (indeclinable)* on the second level. We can also choose one or all possible numbers (*ednina*, *dvojina*, *množina*) (*singular, dual, plural*), for which we choose the declension on the third level, where we enter the given form (*imenovalnik*, *rodilnik*, *dajalnik*, *tožilnik*, *mestnik*, *orodnik*) (*nominative, genitive, dativ, accusative, locative, instrumental*).
- At (3) *pridevniško* (*adjectival*) we have the option *samo_ena_oblika* (*just_one_form*) or *nedoločna_obl* (*indefinite_form*) and *določna_obl (definite_form)* on the second level, depending if there is a distinction between the definite or indefinite form of the adjective, or if there is no distinction. On the same level it is also possible to have elements *primernik* (*positive_form*), *presežnik* (*comparative_form*) and *nesklonljivo (indeclinable)*. With all of them (except the last one) we have to choose the grammatical gender (*m*, *ž*, s) (*masculine, feminine, neutral*) on the third level, the number for each gender on the fourth level (*ednina*, *dvojina*, *množina*) (*singular, dual, plural*), while the fifth level has a set of declensions, where we enter the given form.
- At (4) *števnik (numeral)* on the second level we choose one of the gender possibilities or the element *m_ž_s (m_f_n)*, if the numeral does not indicate the gender. For each gender we then choose the number on the third level, and the declension on the fourth level. With *m_ž_s* (*m_n_f*) we can choose the element *osnovna_oblika_števnika (basic_form_of number)* on the third level, which is then used instead of the nominative and/or accusative and/or the undeclined form, *označitev_osnovne_oblike_števnika (tag_of_basic_form_of_numeral)*, where we indicate the information about whether it is used as a noun, adjective, or its use is irrelevant, and *tudi_kot_nesklonljiv (also_as_indeclinable),* and also other declensions – therein we enter the given form of the numeral.
- At (5) *glagol (verb)* on the second level we choose between: *nedoločnik, namenilnik* (*infinitive, supine*) (they are both end elements), *sedanjik, velelnik* and *del_na_l (indicative, imperative,* and *participle_ending_with_l*). On the third level with the present indicative we choose between the grammatical number, which is, due to its being different from the number in the case of declined words marked as *ednina_glag*, *dvojina_glag*, *množina_glag (singular_verb, dual_verb, plural_verb)*, while each of them has the person category (*oseba_1*, *oseba_2*, *oseba_3) (person_1, person_2, person_3)* on the fourth level. On the third level the case is similar for the imperative form regarding the number category (*ednina_velelnik*, *dvojina_velelnik*, *množina_velelnik*) (*singular_imperative, dual_imperative, plural_imperative).* For dual and plural we must choose the elements *oseba_1* and *oseba_2 (person_1, person_2)*. There is no such division with the singular, where only the 2nd person is possible. Participle

ending with *l* (*del_na_l*) can be of different gender (*m_del*, *ž_del*, *s_del*) (*m_participle, f_participle, n_participle*), where each of them is different in number: *ednina_del*, *dvojina_del* and *množina_del (singular_participle, dual_participle, plural_participle)*.

- At (6) *posamostaljeno* (*nominalized*) we can chose the gender on the second level (the gender is for the positive form), within the gender options on the third level we can choose between the number options, and within them we can chose between cases on the fourth level. On the second level we have the option of choosing between the comparative and the superlative, which contain the same categories as the positive, therefore gender on the third level, number on the fourth level, and declension on the fifth level.
- (7) *izpridevniški_prisl* (izpridevniški prislov) (*deadjectival adverb*) on the second and last level differentiates between the positive form, the comparative form, and the superlative form of the adverb deriving from an adjective (*osnovnik_izpr_prisl*, *primrk_izpr_prisl*, *presež_izpr_prisl*) (*positive_adjectival_adverb, comparative_adjectival_adverb, superlative_adjectival_adverb).*
- (8) *pregibni_povedkovnik (declinable_ predicative)* can take on one of the genders *(m_povedkovnik*, *ž_povedkovnik*, *s_povedkovnik*) *(m_predicative, f_predicative, n_predicative*), these in turn take a different number on the second level *(ed_povedkovnik*, *dv_povedkovnik*, *mn_povedkovnik*) *(singular_predicative, dual_predicative, plural_predicative*), on the second level we can also choose either the comparative or the superlative form *(primrk_izpr_prisl*, *presež_izpr_prisl*) *(comparative_deadjectival_adverb, superlative_deadjectival_adverb*).

In the case where the elements have different names but where the approach is equally explicit and organized, we can enter information on the forms of headwords, depending on which part of speech they belong to.

4.3 Choice of Approach

The complexity of the explicit schema is, similar to the simplicity of the general schema, the outcome of a series of hierarchical decisions made by the lexicographer. If in the example above we perceive six levels of data: (1) adjective > (2) indefinite form > (3) masculine > (4) singular > (5) nominative > (6) entry of given form, then the schema we are creating follows this string of decisions.

Fundamentally, the difference between the two approaches we presented is, for the most part, of a technical nature. Either we want to enter the actual information only as text and, therefore, the elements are designated in a very general manner, because in this case they do not mean anything by themselves except a hierarchical level making it thus essential to have comprehensive lists of possible topics separately for each level (in our case the serial numbers of levels: OZ_1, OZ_2 etc. and a list of allowed topics: seznam1 (*list1*), seznam2 etc.). Or we select the elements for which we already determine not only the hierarchical relation, but also the meaning of the level in an explicit way. A scrutiny of the combinations in the list of options is necessary when choosing the first option since XML Schema itself does not enable it. One option would be to use the *Schematron* standard, which points out the incorrect combinations (e.g. adjective > masculine > 3rd person).

The explicit schema, which requires unambiguously and explicitly tagged information, does not leave lexicographers any doubts. Contrary to the general schema, the drawback of this kind of structure is its complexity; the number of different elements in the structure greatly increases, which makes the schema seem at first sight very complex and branched. This is also the reason why the schema in this article could not be graphically depicted in a transparent way because the size of book format is limited. In spite of these drawbacks, it is a fact[18] that for a lexicographer this type of structure is actually more logical and easier to understand due to the explicitness of the information. A tree structure becomes even more easily manipulated with the right program for lexicographical work, which partially guides the lexicographer when making data entries. A more simple search and data processing of the dictionary database also speaks in favor of the explicit schema, which is important for reviewing already finished work at the time of editing as well as for advanced users of the electronic version. It is also true for linguistic research, where the access to various data in the microstructure is of great importance.

If we initially insisted that both variants of the schema must correspond to the content aspect and did not encounter any obstacles that would deter us from this conviction, we can therefore conclude that, from the point of view of the practical aspect, the suitability of either of the schemas cannot be argued, but, because the making of a dictionary expands over a long period of time, we tend to favor the explicit schema. The technical aspect certainly confirms that it is the best solution.

In our estimation the second option, i.e., the explicit schema of the morphological header, is better for lexicographical work and subsequent usage of the dictionary database because it enables a clear classification of linguistic data into the dictionary database and facilitates search for lexicographical data at a later stage. However, we are aware all along that this solution has certain disadvantages, which prevents it from being ideal.

5 Conclusion

Modern lexicography is inconceivable without the appropriate technical support and the application of software tools. XML, which allows hierarchical structuring of data, has recently been established as the standard format for encoding dictionary

[18] Both schema variants (as well as variants of other parts of the schema structure) were tested among the lexicographic team and it was shown clearly that explicit schema gave better results.

databases and many other language sources. For dictionaries that contain mainly a great number of various data types, the unambiguous identification of data is crucial. The appropriate schema, which exhibits the dictionary structure, provides lexicographers with a clear picture of the dictionary's overall structure, whereas the method of its production affects the difficulty of both the construction and use of the dictionary. Therefore, when designing the schema to be a projection of the dictionary concept, we must firstly consider three aspects, the content aspect, the practical aspect, and the technical aspect, as well as the possibility to use one of the standard formats for encoding lexicographical data. In the Dictionary of the Slovene Literary Language of the 16th Century it is even more crucial to carefully consider the structure of the schema given the fact that the dictionary has a very complex microstructure due to the nature of the material and that fact that it will evolve during a long period of time.

References

Abel, Andrea. 2012. Dictionary Writing Systems and Beyond. In *Electronic Lexicography*, ed. by Sylviane Granger, and Magali Paquot, 83–106. Oxford: Oxford University Press.

Ahačič, Kozma, Andreja Legan Ravnikar, Majda Merše, Jožica Narat, and France Novak. 2011. *Besedje slovenskega knjižnega jezika 16. stoletja*. Ljubljana: Založba ZRC, ZRC SAZU.

Berglund, Anders. 2006. *Extensible Stylesheet Language (XSL) Version 1.1.* <http://www.w3.org/TR/2006/REC-xsl11-20061205/>.[19]

Birbeck, Mark, Markus Gylling, Shane McCarron, and Steven Pemberton. 2010. *XHTML™ 2.0.* <http://www.w3.org/TR/2010/NOTE-xhtml2-20101216/>.

Bray, Tim, Jean Paouli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau, and John Cowan. 2006. *Extensible Markup Language (XML) 1.1 (Second Edition).* <http://www.w3.org/TR/2006/REC-xml11-20060816/>.

Clear, Jeremy. 1987. Computing: Overview of the Role of Computing in Cobuild. In *Looking Up: An Account of the COBUILD Project in Lexical Computing,* ed. by John M. Sinclair, 41–61. London, Glasgow: Collins ELT.

Erjavec, Tomaž, Ines Jerele, and Maša Kodrič. 2011. *Izdelava korpusa starejših slovenskih besedil v okviru projekta IMPACT.* In *Meddisciplinarnost v slovenistiki. Obdobja 30*, ed. by Simona Kranjc, 41–47. Ljubljana: Znanstvena založba Filozofske fakultete.

Hunter, David, Jeff Rafter, Joe Fawcett, Eric van der Vlist, Danny Ayers, Jon Duckett, Andrew Watt, and Linda McKinnon. 2007. *Beginning XML*. Indianapolis: Wiley Publishing.

Krek, Simon. 2004. Slovarji serije COBUILD in formalizacija definicijskega jezika. *Jezik in slovstvo* 49/2: 3–16.

Merše, Majda, France Novak, and Francka Premk. 2001. *Slovar jezika slovenskih protestantskih piscev 16. stoletja : poskusni snopič*. Ljubljana: Založba ZRC, ZRC SAZU.

Smrž, Pavel, 2001. Slovníková data ve formátu XML. In *Slovenčina a čeština v počítačovom spracovaní. Zborník referátov zo seminára. Bratislava 26. – 27. októbra 2001*, ed. by Aleksandra Jarošová, 175–187. Bratislava: Veda.

[19] All referenced websites in the article were accessible on 9 June 2012.

Thompson, Henry S., David Beech, Murray Maloney, and Noah Mendelsohn. 2004. *XML Schema Part 1: Structures: W3C Recommendation 28 October 2004*. <http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/>.

ABBYY Lingvo Content <http://www.abbyy.com/lingvo_content/>.

iLEX <http://www.emp.dk/>.

IDM DPS <http://www.idm.fr/products/dictionary_writing_system_dps/27/>.

TshwaneLex <http://tshwanedje.com/tshwanelex/>.

Termania <http://www.termania.net>.

Prispelo januarja 2013, sprejeto marca 2013
Received January 2013, accepted March 2013

Oblikovanje XML-sheme za leksikografske projekte na primeru Slovarja slovenskega knjižnega jezika 16. stoletja

Prispevek prikazuje in pojasnjuje, kako je bila oblikovana struktura oblikoslovnega zaglavja Slovarja slovenskega knjižnega jezika 16. stoletja, zapisana v računalniškem formatu XML-shema, skladno s konceptualnimi zahtevami sestavljavcev slovarja. Izdelani in na podlagi različnih meril analizirani sta bili dve v osnovi bistveno različni shemi oblikoslovnega zaglavja, tj. posplošena in eksplicitna shema.

Da si sodobne leksikografije ni mogoče predstavljati brez ustrezne računalniške podpore in uporabe programskih orodij, je danes samoumevno. Slovarska baza za Slovar slovenskega knjižnega jezika 16. stoletja bo nastajala v standardnem formatu XML, ki omogoča večnivojsko hierarhično strukturirane podatkovne baze (drevesna struktura), povezovanje s sklici, zahtevno iskanje in različne obdelave podatkov. Ker so datoteke XML v resnici navadne besedilne datoteke, so prenosljive med različnimi programi in operacijskimi sistemi, kar je v času stalnega napredka informacijske tehnologije velikega pomena pri dolgoročnem delu, saj lahko slovarsko podatkovno bazo ne glede na uporabljeni program za izdelavo kasneje uporabimo v katerem koli programu, ki zna brati navadne besedilne datoteke.

Slovarska struktura Slovarja slovenskega knjižnega jezika 16. stoletja je izdelana v formatu XML-shema. Ta med drugim določa, kateri elementi so v slovarski bazi dovoljeni, kakšna so hierarhična razmerja med njimi in kakšen je njihov vrstni red, kakšne so možnosti njihovega kombiniranja oziroma izključevanja in kolikokrat se določen element lahko ponovi, kadar želimo navesti več zaporednih enakih elementov. Ker leksikografska programska orodja izkoriščajo XML-shemo za pomoč leksikografom pri izdelavi pravilne strukture geselskih člankov, je bilo pri izdelavi sheme za oblikoslovno zaglavje poleg vsebinskega vidika smiselno upoštevati tudi praktični in tehnični vidik. Od njiju je namreč odvisna hitrost izdelave slovarja, pa tudi (ne)zapletenost kasnejše uporabe slovarskih podatkov za jezikoslovne raziskave in uporabo v drugih jezikovnih opisih.

Ker smo izhajali iz tega, da morata obe predlagani shemi ustrezati leksikografskemu vidiku, sta bila za izbiro med dvema različnima shemama odločilna praktični in tehnični vidik. Na podlagi analize prednosti uporabe ene ali druge različice sheme je bila zaradi jasne strukturiranosti podatkov in enostavnejše izdelave slovarja izbrana eksplicitna shema, ki daje slovarski podatkovni bazi večjo vrednost tudi za kasnejšo uporabo slovarskih podatkov.

Creating XML Schemas for Lexicographical Projects in the Case of the Dictionary of the Slovene Literary Language of the 16th Century

This paper depicts and explains how the structure of the morphological header of the Dictionary of the Slovene Literary Language of the 16th Century was written in electronic XML Schema definition language, in accordance with the lexicographers' conceptual requirements. Two substantially different schemas, i.e., a general and an explicit schema of a morphological header, were designed and analyzed on the basis of different criteria.

It is indeed inconceivable to imagine modern lexicography today without the appropriate technical support and software tools. The Dictionary of the Slovene Literary Language of the 16th Century database will be created in standard XML format, thus enabling a hierarchically structured database (tree structure), references, advanced search, and a variety of data processing. XML files are in fact nothing more than plain text files, transferable between different programs and operating systems. In a time of continual progress and change in information technology this is extremely significant for long-term work considering that the dictionary database, regardless of the program used, can later be applied to any program that is able to read plain text files.

The structure of the Dictionary of the Slovene Literary Language of the 16th Century is designed in XML Schema definition language, which inter alia defines the elements allowed in the dictionary database, the hierarchical relations between them and their order, the possibilities of their combination or exclusion, and the number of times a certain element can occur, when we wish to specify more than one consecutive element. Since the lexicographical software tools exploit XML Schema in order to support lexicographers in creating the correct structure of the dictionary entries, it was crucial to take into consideration the practical and technical aspects in addition to the content aspect for the process of making the schema for the morphological header. These two additional aspects contribute to the speed of the dictionary production and the complicatedness, or lack thereof, of subsequent use of dictionary data for linguistic research purposes, as well as its use in other language descriptions.

We proceeded from the fact that both presented schemas meet the requirements of the content aspect, which left the practical and technical aspects to be the decisive elements in our choice of schema. Based on the analysis of the advantages when using one or the other version of the schema, it became clear that, due to the transparent structure of data and easier construction process of the dictionary, the explicit schema was our preferred choice; moreover, it gives the dictionary even greater value for subsequent use of lexicographical data.