Professional paper

Slovenian Biographical Lexicon – From a Digital Edition to an On-Line Application

Jan Jona Javoršek
Department of Experimental Particle Physics
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
jan.javorsek@ijs.si

Petra Vide Ogrin
Slovenian Academy of Sciences and Arts, Library
Novi trg 3, Ljubljana, Slovenia
petra.vide@zrc-sazu.si

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Summary

We presents the digitization of Slovenian Biographical Lexicon (SBL), an extensive publication that used bio-bibliographical methods to provide synthetic assessments of work and significance of historical figures on the basis of primary sources. SBL has been out of print for a long time, but the publication has been seen as an important resource for encyclopaedic and reference editions and research in the Slovenian humanities, social sciences and history of the natural sciences. Therefore, Slovenian Academy of Sciences and Arts (SASA) and the Scientific Research Centre of the SASA decided to produce a freely available on-line digital re-edition of SBL. In the process of digitalization, manually corrected OCR has been semi-automatically converted to XML-based Text Encoding Initiative format (TEI P5). Its extensive annotation vocabulary, notably from the biographical and prosopographical modules, has been used to mark-up as much data as possible. The resulting XML document has become the data resource of an online digital repository based on Fedora Commons platform, where we implemented an infrastructure of XML processing methods on top of native relationships and a Lucene/SOLR based search engine to produce a full-fledged web application and search engine with browser, metadata and web application interfaces.

Key words: digital library, encyclopaedia, TEI-XML markup, document repository, Fedora Commons, XSLT workflows, search, Lucene, SOLR

Introduction

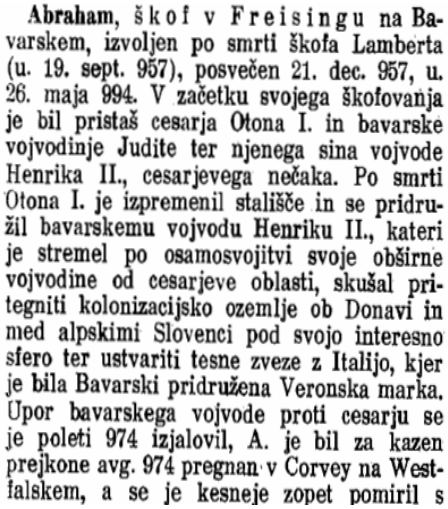
The Slovenian Bibliographical Lexicon (SBL, 1925-1991) has been conceived as a publication that was to give an accurate picture of Slovenia’s cultural life, from its beginnings up to the contemporary time by including everybody who participated in the cultural development, either of Slovenian origin, born on Slovenian soil or influencing Slovenian cultural life.

This broad aim resulted in the proposed list of 2,335 names, mostly from the fields of humanities and social sciences, proposed by the original editorial board. In its long history, this list has been changed and expanded: at first due to the need to include other disciplines and areas, such as the natural sciences, later due to changes in perspective, especially the WW2, when participants in recent had to be included and the focus has been shifted to include “increasing development of natural sciences, modern technologies and their applications, as part of the spiritual superstructure” (SBL, vol. 15, 1991). In spite of several eliminations from the original list, the final publication in 1991 comprised as many as 5,031 biographical entries, with more than 5,100 persons covered. Since the publication was published sequentially, it is important to consider that the criteria for different published volumes have therefore varied significantly.

The second aim of the publication was to be both informative and exhaustive, so much substantial information had to be included in rather short articles (with several longer exceptions) and it has been decided that the data in the articles had to be been checked against relevant historical materials and pre-existing publications., e.g. biographical and other dates are always compared to dates in registers and other primary documents, literary citations are compared with originals, sources are cited at the end of the articles and the publication includes an index of all person names that appear in the articles and a list of abbreviations.

As a result, SBL contains a surprising amount of high-quality information and references and remains to this day a precious resource for encyclopaedic and reference editions and research in the Slovenian humanities, social sciences and history of the natural sciences. It had, however, two severe drawbacks: the original edition has quickly become difficult to find and the information, once published, has never been updated, so much care has to be taken to consider the time of publication of each article. The present project of digitalization of SBL has been started by the Slovenian Academy of Sciences and Arts (SASA) and the Scientific Research Centre of the SASA to make this precious resource available again, this time in the form of a freely accessible on-line edition, and has been based on previous similar projects (cf. Erjavec, Ogrin, 2005). We aim to describe here the steps taken on the path from the original publication towards a fully interactive, searchable and cross-indexed on-line edition. The first

steps of the process, from digitalization using OCR and manual revision to semi-automatic encoding and mark-up in the form of Text Encoding Initiative XML document (TEI P5), will be summarized (see Ogrin, Erjavec, 2007 for a more detailed treatment of this topic) before we consider the methodology and implementation of and on-line digital repository for the digital edition that can function as a flexible web application. We will present several possibilities offered by our implementation, some of which remain for further experiments. In closing, we wish to present some possible solutions to the hardest problems encountered in this work, from normalization of data and abbreviation expansion to the question of treatment of the original in view of the necessities of every day users who wish to have access to updated information. It is worth noting at this point that the digital edition in current form is available for testing and the reader can follow the presentation at the following URL: http://nl.ijs.si/fedora/sbl/.



Abraham, škof v Freisingu na Bavarskem, izvoljen po smrti škofa Lamberta (u. 19. sept. 957), posvečen 21. dec. 957, u. 26. maja 994. V začetku svojega škofovanja je bil pristaš cesarja Otona I. in bavarske vojvodinje Judite ter njenega sina vojvode Henrika II., cesarjevega nečaka. Po smrti Otona I. je izpremenil stališče in se pridružil bavarskemu vojvodu Henriku II., kateri je stremel po osamosvojitvi svoje obširne vojvodine od cesarjeve oblasti, skušal pritegniti kolonizacijsko ozemlje ob Donavi in med alpskimi Slovenci pod svojo interesno sfero ter ustvariti tesne zveze z Italijo, kjer je bila Bavarski pridružena Veronska marka. Upor bavarskega vojvode proti cesarju se je poleti 974 izjalovil, A. je bil za kazen prejkone avg. 974 pregnan v Corvey na Westfalskem, a se je kesneje zopet pomiril s

```
<div>
  <listPerson>
    <person n="main">
      <sex value="1"/>
      <persName>
        <name>Abraham</name>
        <roleName type="eccl">škof</roleName>
      </persName>
      <occupation>duhovnik</occupation>
      <death>
        <date when="0994-05-26">26. maja 994</date>
      </death>
    </person>
    <person n="author">
      <sex value="1"/>
      <persName key="M. Kos.">
        <surname>Kos</surname>
        <forename>Milko</forename>
      </persName>
    </person>
  </listPerson>
  <p>Abraham, škof v Freisingu na Bavarskem, izvoljen
po smrti škofa Lamberta (u. 19. sept. 957), posvečen 21.
```

Figure 1: An SBL article excerpt with its TEI XML encoding

Text Encoding Initiative: the Data Source

Encoding of SBL is based on open standards and software, in particular the Text Encoding Initiative P5 Guidelines, since TEI has been used as the encoding standard for the digitalization of the Slovenian Bibliographic Lexicon.

TEI produced recommendations or guidelines for the creation and processing of electronic texts for better interchange and integration of scholarly textual data in all languages and from all periods (Burnard, 1988). We used the latest edition of TEI Guidelines, TEI P5, finalized in 2007 (Burnard, Bauman, 2007), since it provides important new encoding features, including new support for manuscript descriptions, multimedia and graphics, stand-off annotation, representa-

tion of data pertaining to people and places and improved specifications for encoding textual alternatives. Additionally, TEI P5 takes advantage of the power of XML schema languages, so that other XML tag-sets, such as MathML or SVG, can now be referenced from within a TEI document and a TEI document can be embedded within other types of XML documents, such as METS and MODS records (Burnard, Bauman, 2007), which turned out to be crucial for the implementation of our on-line repository, since this makes TEI a well-behaved XML citizen, able to take part in any, however complex, XML processing chains and composite documents.

XML Data Source Structure

The vast majority of SBL articles present information on the life and actions of a single person, while some describe well known families, detailing life and work of several members of the family. An article usually starts with the name of the person or the family, its variants, mostly those used towards the end of their life or the most generally used, followed by a chronological summary of the person’s life and activity, including birth, death, locations, occupations, activities etc. An article may consist of one (usually) or many paragraphs, depending on the exhaustiveness of the article, and is written in dense language, using abbreviations wherever possible, ending with a brief bibliography and other materials relevant to the person, such as portraits or photographs.

The text of the articles has been digitized and manually revised to fix OCR errors before it was automatically converted into the basic TEI-XML format. In the next stage, segments of text that needed to be marked up but could not be identified automatically had to be tagged manually, in particular with details such as different variants of names (linguistic and orographic variations, married names, ecclesiastic names and titles, pseudonyms, complex name parts in the case of foreign names and names with denotation of nobility etc.), making the process slow and error-prone. Since the original data was not normalized, considerable effort had to be spent to achieve high quality TEI XML mark up, and some work with data normalization is still ongoing. (The major aspects of this conversion process have been reported in more detail in Vide Ogrin, Erjavec 2007). In this manner, essential information about the subject of the article and its bibliographical section have been encoded with special purpose elements from TEI P5 biographical and prosopographical modules, representing each article as a element starting with a <listPerson> element, which contains the detailed information on the subject of the article (names, sex, birth/death date and location, locations of activities, occupations or activities etc.), but also meta data, (volume and year of the first publication, author of the article, revision status etc.). This element is followed with one or more <p> elements with the article text and a <listBibl> element with the extracted bibliographical data.

Obviously, the actual structure to a certain extent depends on the information of that particular article, and so the type and number and elements varies considerably (ie. marriage, ordination, exile, further education, number of occupations, residence, active period etc.). This makes any mapping into a more formal structure, i.e. a relational database, at least awkward.
There is a number of further details that could be extracted from the text but meticulous manual intervention would be required to achieve suitable accuracy. The most important of these are activities undergone by the person, encodable in the <occupation> tag or tags, and locations and times of these activities, encodable by the <floruit> tag.

Anatomy of an XML-based Document Repository

We have evaluated a number of possible platforms to serve as the base of an on-line web edition of SBL with integrated query and search tools. One possibility considered was PhiloLogic[1], a full-text search, analysis, and retrieval tool developed by the ARTFL Project[2] and the Digital Library Development Center at the University of Chicago. PhiloLogic uses an abstract representation of document structure, projecting the XML data into sets of related database tables, so that the application can search document structure and refine word searching by using the XML structure (Cooney et al., 2007). However, in the end we have choosen the Fedora Commons platform (see The Fedora Project[3]), an extensible framework for storage, management and dissemination of complex objects and object relationships implemented as a portable Java web application (Lagoze et al, 2006). Fedora Commons became the repository on top of which we created a digital library of bibliographical articles with browsing, searching and querying interfaces: a digital library that is presented as an on-line web service and application.
Fedora Commons represents its digital objects as a collection of data streams, where each document is specified as an XML document (Fedora Commons has a native format, called FOXML, but also supports METS). Data streams can be of different types: formally, they can be created as embedded XML documents, as managed independent files in the repository (used for binary files, i.e. images, PDF documents and similar) or as external URI-specified documents. In addition, each object has a number of infrastructure-supporting data streams, all in the form of embedded XML documents. Among them, there is a Dublin Core data stream to contain object meta-data, a RDF data stream to declare inter-object relationships, and internal FOXML revision specifications to allow tracking of object history.

[1] http://philologic.uchicago.edu/

[2] http://humanities.uchicago.edu/orgs/ARTFL/

[3] http://fedora-commons.org/

Furthermore, Fedora Commons objects have dissemination methods (analogues to object or class methods in object-oriented systems), implemented as web application interfaces to objects and their contents (both REST and SOAP interfaces are supported). Since version 3 of the platform, a new Content Model Architecture has been introduced under which dissemination methods are specified with three special objects types: Content Model objects specify available methods and necessary data streams for the dissemination methods they declare, Service Definition objects define a web API for dissemination methods and Service Deployment objects use WSDL (web service definition language) to specify the actual web application API calls necessary to execute a dissemination method request (cf. Fedora Commons Content Model Architecture documentation for version 3[4]).

All the required information, such as the necessary data for each dissemination method call, supported data-types and the manner of invocation to produce the result, are specified in the form of embedded XML documents in the three types of objects, which are otherwise structured as any other object in the repository. To add dissemination methods from one or several different Content Models, it is therefore sufficient to add a special relationship to an object, refering to the Content Model in question. Its dissemination methods will become available under the access URI of the object, contained in a path element of of the Content Model’s name. Furthermore, even the core Fedora Commons features, such as object introspection and direct data stream access, are implemented in this way using the default Content Model.

This extensible and standards-based Content Model Architecture in combination with a number of web services, namely the SAXON XSLT processor, an image manipulation library, a RDF query interface to the object relationship RDF store, a simple search interface to Dublin Core meta-data and object properties and the Fedora Generic Search interface to a number of optional search engines, provide an infrastructure for development of rich application interfaces and complex multi-layered digital repositories using standard technologies and XML workflows.

From XML Datasources and Workflows to an Online Application

In the Fedora Commons framework, each dissemination method is realized as a web application call (using REST of SOAP methods) with a number of arguments, usually one or several of the object’s data streams. But a data stream can have the form of an URI-specified data stream, referin to another dissemination method and thus resulting in a chain of processing calls. While this approach can be used with binary data, i.e. to apply a number of transformations to an image, it is usually used to create an XML workflow. Such XML workflows

[4] http://fedora-commons.org/confluence/display/FCR30/Content+Model+Architecture

have become the backbone of our application since they allow us to poll together XML data from several sources, such as object data streams and object relationship query results, to form the final XML response, usually in the form of an XHTML rendering in the users browser.
Essentially, there are two kind of data objects in our application: collections and articles. Collections use inter-object relationships to represent different views of the data to the user: these are the top-level object, containing all the other views, letter-objects and volume-objects that allow browsing alphabetically through lists of articles or browsing through the articles in the units of their publications, and search-result objects that take a search query and represent its results as a collection of objects.
Articles are much simpler – in essence they transform the TEI data to an XHTML-encoded web page. However, the resulting page contains a number of context-dependent links, including facilities such as a search interface, browsing links (previous, next), links to instant search queries that provide article lists representing i.e. members of the same occupation class, members of the same generation or all the contemporaries of the subject of the article, but there are also other links, such as a link to all the articles by the same author, accessed via the author name, etc.

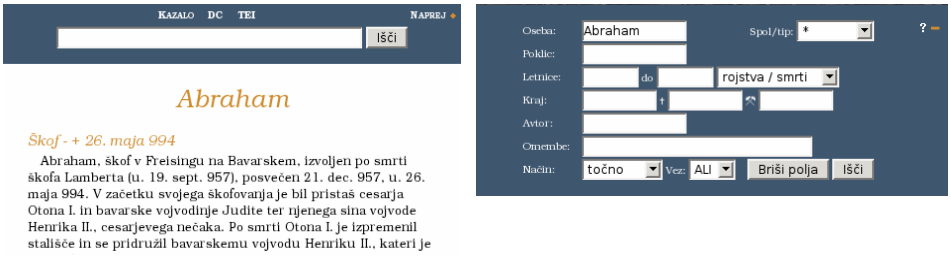


Figure 3: A SBL Article rendering next to the advanced SBL search form

At the same time, all the objects provide direct links to their TEI XML source and their meta data in Dublin Core encoding, transformed from the TEI header-like structures in <listPerson> element of each article (cf. Miller, Brickley 2001 and Liddy et al., 2002). This seemingly trivial feature is actually essential: it means that it is possible, through a public API, to access all of the original TEI document data and all of the meta data, structured in a standard-compliant way. With the combination with platform-integrated support for the Open Archives Initiative Metadata Harvesting Protocol (OAI-MHP), this makes our digital repository very easy to integrate with other systems (cf. Benjamin and Siberski, 2002, Ward 2004).
There is a number of simpler objects, mostly renderings of parts of TEI header information, that simply convey meta data about the whole collection in an accessible form – but they all obey the same logic and use the same infrastructure.

This architecture, in spite of its relative simplicity, has allowed us to construct a flexible and efficient user interface. In general, all the context information has become clickable or otherwise accessible through simple links, making the browsing interface extremely powerful. But the true power of the implementation comes from its search interface.

Search and Query Interface

Fedora Commons provides a simple integrated search system, capable of simple searches on Dublin Core metada ta and object properties, but a much more powerful system, Fedora Generic Search, is available. The power of this system derives from the fact that is simply provides native Fedora Commons interfaces between an external search system and Fedora Commons API.

In our application, we have chosen to implement Fedora Generic Search on top of SOLR, a search system based on Apache Lucene search and indexing library. In this set-up, Lucene library can use Fedora Commons API to index the document contents, using specially crafted rules (in the form of XSLT stylesheets) to break up the documents in a number of searchable text fields, and the repository gains a search interface with full power of Lucene query language (cf. Hatcher, Gospodnetic 2004), while SOLR takes over the interface and formatting of query results in an easy-to parse XML list.

The power of the interface has been most useful while crafting special queries, such as the context links for different SBL article features, but due to the complexity involved in the use of the flexible Lucene query language, this is hardly the optimal solution for the average user of the system.

To solve this problem, we have implemented two use search interfaces: the simple search is targeted by default at the most often requested fields, namely the subjects’ names, places of their birth and death, and their occupation descriptions. This interface in fact accepts the full Lucene syntax, so it can be used both for simple searches by general users and for complex queries by advanced users.

The secondary interface can be accessed by a click on an expansion button. It presents a form with several fields that easily enable an average user to compose even fairly complex queries, selecting different indexed fields and even using advanced features such as full-text searches, proximity searches, number or date ranges, fuzzy searches etc. This interface is implemented by a secondary script that parses the form and converts its data into a single Lucene query string. Our initial testing with users has been reasonably successful: it makes it trivial to find, for example, female writers, who lived and worked in the period between 1830-1860 in Ljubljana, or priests, who were also philosophers and born in Maribor etc.

The same system can be used for experimental research on the data, especially if one wants to analyse the particularities of the original SBL publication. It is now easy to compare, for example, the average length of articles with the num-

ber of articles contributed by an author (showing the difference between regular contributors and specialists for narrow fields), to plot the average length of life by year of birth or by occupation (and find extreme cases, such as ‘ exceptionally short life spans for heroes of WW2 and revolution). In fact, a number of such queries with graphic interfaces, together with usual on-line features, for example a listing births and deaths on the current date, are being included in the current version of the application. Obviously, this opens up possibilities for further research that falls outside the scope of the present paper.

Conclusions

The project of the digital edition of SBL has reached production stage, where a complete digital edition is hosted in an on-line digital library and an on-line user interface is available. The main goal of the project has been been achieved: the valuable reference information captured in the lexicon is now again available to the research community and general public, this time enhanced with cross-linking, context information and an advanced query and retrieval system that facilitates it use.

But there are several objectives that are still being worked on. Primarily, we would like to extract and encode further extents of information, since we now have a functional framework thank enables us to use the information, once marked up, to provide further features. The foremost points are the two crucial pieces of information: the subject’s name and occupation, representing the two identificators used to mos often find subjects of interest. The work of manual annoation of name parts with further corrections in spelling and markup of names is being finished at the time of this writing. At the same time, we are working on normalization of occupation specifications (there are more than 1500 different strings for occupations or activities in the SBL articles) and introducing a simple taxonomy to enable meaningful grouping and searching.

Furthermore, we have plans to normalize names of places and annotate them with geographic identifiers. Since we also want to mark-up any names mentioned in the article texts, we are planning to develop a Named Entity Recognition (NER) tool, (Jackson, Moulinier, 2002; Bekavac, 2002) for Slovenian, and we have gathered substantial databases of names and places for this purpose. This task is further complicated by the fact that also Slovenian inflected forms need to be normalised, and the many abbreviated entities in the texts have to be disambiguated, expanded and properly inflected.

In closing, we are happy to report that our system will be used by ongoing and new project in the field of bibliographic publications in Slovenia and will likely become the platform for the digital publication of Slovenian Biographical Lexicon 2 and for the Slovenian Bibliographical Hub which is to integrate most available resources in this domain.

References

Bekavac, Božo. Strojno obilježavanje hrvatskih tekstova – stanje i perspektive // *Suvremena lingvistika.* 53-54 (2002), 173-182. http://nl.ijs.si/et/tmp/matija/BB_disertacija_verT9.pdf (2007-09-14)

Benjamin, Wolfgang Nejdl, Siberski, Wolf , 2002: OAI-P2P: A Peer-to-Peer Network for Open Archives, v Workshop on Distributed Computing Architectures for Digital Libraries - ICPP2002.

Burnard, Lou. Encoding standards for the electronic edition // Znanstvene izdaje in elektronski medij: razprave / Ogrin, M. (ed.). Ljubljana: Založba ZRC, ZRC SAZU, 2005, 25-42

Burnard, Lou. Report of Workshop on Text Encoding Guidelines // *Literary & Linguistic Computing*. 3 (1988), 131-133

Burnard, Lou; Bauman, Syd. TEI P5:Guidelines for Electronic Text Encoding and Interchange (TEI P5). Text Encoding Initiative Consortium. HTML Version. Oxford, 2007. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ (2009-08-19)

Cankar, Izidor et al. (eds). Slovenski biografski leksikon. Ljubljana: Slovenska akademija znanosti in umetnosti, 1925-1991

Cooney, Charles M. et al.. Extending PhiloLogic. April 2007.

http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=175 (2007-09-09)

Erjavec, Tomaž; Ogrin, Matija. Digital Critical Editions of Slovenian Literature: an Application of Collaborative Work Using Open Standards. // *From Author to Reader: Challenges for the Digital Content Chain: proceedings of the 9th ICCC International Conference on Electronic* Publishing, Arenberg Castle / Dobreva, M.; Engelen, J. (eds.). Leuven: Peeters, 2005, 151-156

Hatcher, Erik, Gospodnetic, Otis 2004: Lucene in Action. New York: Manning Publications.

Jackson, Peter; Moulinier, Isabelle. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. Amsterdam: John Benjamins, 2002, 180-185

Kahn, Robert; Wilensky, Robert. A framework for distributed digital object services. Working Paper, 1995. cnri.dlib/tn95-01, http:// www.cnri.reston.va.us/k-w.html

Lagoze, Carl; Payette, Sandy; Shin, Edwin; Wilper, Chris. “Fedora: an architecture for complex objects and their relationships”, *International Journal on Digital Libraries*, vol. V6, no. 2, 2006, 124-138.

Liddy, Elizabeth D., Allen, Eileen, Harwell, Sarah, Corieri, Susan, Yilmazel, Ozgur, Ozgencil, Ercan N., Diekema, Anne, Mccracken, Nancy, Silverstein, Joanne in Sutton, Joanne, 2002: Automatic metadata generation & evaluation. V SIGIR ‘02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM Press, 401-402.

Miller, Eric, Brickley, Dan, 2001: Expressing Simple Dublin Core in RDF/XML, Dublin Core Metadata Initiative Proposed Recommendation.

The Fedora Project: An Open-Source Digital Repository Management System, http://fedora-commons.org/

Vide Ogrin, Petra; Erjavec, Tomaž: Towards a Digital Edition of the Slovenian Biographical Lexicon. In: *The Future of Information Sciences: Digital Information and Heritage : Proceedings of the 1st International Conference The Future of Information Sciences - INFuture 2007* / Seljan, Sanja ; Stančić, Hrvoje (ur.). Zagreb : Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu, 2007. 115-124.

Ward, Jewel, 2004: Unqualified Dublin Core usage in OAI-PMH data providers. OCLC Systems & Services, 20(1), 40-47.