# Slovenian Biography

**Tomaž Erjavec[1], Joh Dokler[2], and Petra Vide Ogrin[3]**
[1]Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
[2]Seven Past Nine Ltd.
Vrtača 3, Ljubljana, Slovenia
[3]Slovenian Academy of Sciences and Arts, Library
Novi trg 3, Ljubljana, Slovenia
E-mail: tomaz.erjavec@ijs.si, joh.dokler@gmail.com, petra.vide@zrc-sazu.si

**Abstract**

The paper presents the Slovenian Biography portal and data. The Slovenian Biography currently comprises almost 9,000 person entries, which are composed of the free text from three source biographical lexicons and semi-automatically extracted structured authority records. These contain detailed information about the person name(s), date and geolocated place of birth and death, as well as the occupation(s) of the person from a rich developed taxonomy, covering both historical and contemporary occupations. The encoding of the data follows the Text Encoding Initiative Guidelines, in particular its module for biographical and prosopographical data. The Web portal of the Slovenian Biography is built on the BaseX XML database engine, queried by XPath / XQuery expressions, the Lucene search engine and the Django web framework. The portal allows faceted search, visualisation on a map, export of the entries in TEI etc. The SB is still work in progress, with new entries being added on regular basis. We are also working on the enrichment of its encoding, e.g. marking up all person names appearing in the text, adding relations between persons and expanding the numerous abbreviations, carried over from the print editions.

**Keywords:** Slovenian biography, TEI encoding, Structured data, Web portal, Natural language processing



## 1. Introduction

Slovenia only became an independent country in 1991, however, long before that Slovenians had a very strong national identity, mostly centred on their language, so Slovenians, i.e. people born where Slovenian is or was spoken, or having significant influence in these areas, can be traced back for over one thousand years.

This paper presents the data and portal of the Slovenian Biography[1] (SB), detailing the life and work of important Slovenians. It comprises data from three biographical lexicons, the first two having been published in print in several volumes:

- *SBL*: the Slovenian Biographical Lexicon (Cankar et al., 1925–1991), containing 5,048 entries;
- *PSBL*: the Primorska Slovenian Biographical Lexicon (Jevnikar et al., 1974–1994), containing 4,429 entries;
- *NSBL*: the New Slovenian Biographical Lexicon (Svetina et al., 2013–), currently containing 455 entries.

*SBL* is included in the SB in its entirety; *PSBL* is still in the process of digitisation, with currently about 40% of the entries included, while *NSBL* is being added to the SB gradually as its publication is an ongoing process.

The *SBL*, the largest of the three lexicons, strives to be an authoritative resource: the authors of the articles follow strict scientific standards, using a responsible historical and biographical method, meaning that all data is checked against the relevant historical materials and pre-existing publications. For example, the biographical and other dates are always compared to those in registers and other primary sources, 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. It is thus a reference work and a precious resource for any serious research in the fields of Slovenian humanities, social sciences and history of natural sciences.

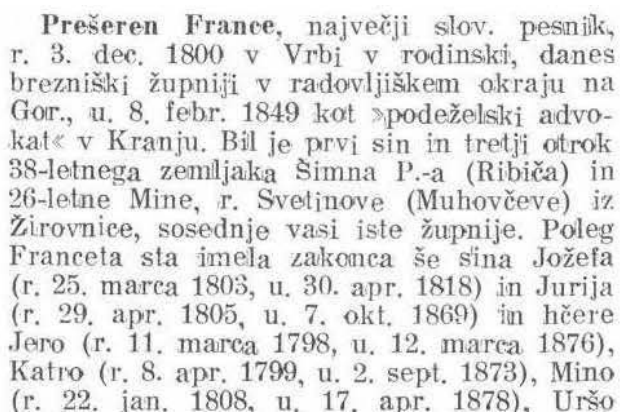
The SB is a long-term joint project of the Slovenian Academy of Sciences and Arts (Slovenska akademija znanosti in umetnosti – SAZU), the Scientific Research Centre at SAZU, the Jožef Stefan Institute and Seven Past Nine Ltd. This paper details its current state: Section 2 overviews the digitisation and up-translation of the source biographical lexicons to XML, including the editing environment and the structure and content of the resulting documents, Section 3 explains the technologies used and the user interface of the web portal, Section 4 overviews our on-going work, and Section 5 gives some conclusions.

## 2. Developing the database

The digitisation of the printed edition of *SBL* (the beginning of an example entry is given in Figure 1), comprising 16 volumes, began in 2007, in order to make this main biographical resource freely accessible. The digitised version was made available on-line, using a platform based on Fedora Commons (Javoršek et al., 2009a, 2009b), since discontinued, as we migrated to the current platform presented in this paper.

After finishing the *SBL*, the work moved on to digitising the *PSBL*; the scanning, OCR and manual correction of the transcription has now also been completed, and its articles are being added gradually to SB, where the extracted structured data is still being manually corrected before the articles are fully integrated into the SB.

[1] http://www.slovenska-biografija.si/



**Prešeren France**, največji slov. pesnik, r. 3. dec. 1800 v Vrbi v rodinski, danes brezniški župniji v radovljiškem okraju na Gor., u. 8. febr. 1849 kot »podeželski advokat« v Kranju. Bil je prvi sin in tretji otrok 38-letnega zemljaka Šimna P.-a (Ribiča) in 26-letne Mine, r. Svetinove (Muhovčeve) iz Žirovnice, sosednje vasi iste župnije. Poleg Franceta sta imela zakonca še sina Jožefa (r. 25. marca 1803, u. 30. apr. 1818) in Jurija (r. 29. apr. 1805, u. 7. okt. 1869) in hčere Jero (r. 11. marca 1798, u. 12. marca 1876), Katro (r. 8. apr. 1799, u. 2. sept. 1873), Mino (r. 22. jan. 1808, u. 17. apr. 1878). Uršo

Figure 1: The beginning of an entry in the printed edition of the SBL.

Finally, the publication of *NSBL* is an ongoing process and volumes are to be published successively in alphabetical order. The articles are born-digital and the structured biographical data is being added manually and integrated into the SB simultaneously with the free text articles, as they become available.

### 2.1 Editing environment

The SB uses XML as a markup language to encode and structure its information. The XML is compliant with the Text Encoding Initiative[2] Guidelines for Electronic Text Encoding and Interchange (TEI, The TEI Consortium), which is an extensive and flexible schema used to represent texts in digital form. We chose TEI as our base encoding not only because of the wide range of text types that it covers and its continuous development, but also because we have substantial prior experience in using it for developing e.g. text-critical digital editions of Slovenian literature (Erjavec & Ogrin, 2005).

The editorial team works directly on the XML using the specialized XML editor Oxygen that supports data validation based on the TEI schema. This allows editors to encode rich data structures in a data-safe manner through the use of schema validation without the need to implement complicated user interfaces. In this sense, XML is a user interface and one that editors can extend themselves. In the context of the project, we did in fact experiment with a simplified web based user interface (on top of the relational database models) that provided editing without prior knowledge of XML. However, this proved to be both limiting and inflexible in terms of project growth. But more importantly, editors, once they learned XML, preferred working with XML.

A very important component of our XML workflow is the use of Git as the version control system. This allows different editors to simultaneously work on the same source XML files, track changes and history, and resolve conflicts. Since Git can be a challenging technology to use in its original form (using the command line) we use a GUI offered by GitHub[3]. While this version provides only a subset of Git functionality it proved to be adequate for our use case and importantly, required the editors to learn only a handful of concepts and commands like pull, commit and push.

### 2.2 Database design and content

The complete database of the SB consists of four TEI documents. *SBL, PSBL* and *NSBL* are each encoded as one TEI document, which contains the TEI header, giving its metadata, and the body with full text entries. The entries are lightly structured, containing a series of paragraphs and the bibliography list. Crucially, each entry has a reference to the person entry in the fourth "authority" (or "index") document (*SBI*), which contains the structured biographic data, further discussed below.

Table 1 gives a quantitative overview of the current state of the database. The first three lines give the number of full-text entries in the *SBL*, *PSBL*, and *NSBL* documents, and the fourth their sum. Next, the numbers of entries in the authority *SBI* document are given. The *SBI-family* is the number of family entries (as opposed to individual persons) included in the document. Next, *SBI-main*, the number of "main" person entries is given – these are the entries that are linked to from the individual lexicons. It should be noted that the sum of the entries in the lexicons is greater than the sum of the family and person records in the authority document, as some persons from the three lexicons overlap. The table next gives the number of *SBI-sub* entries, i.e. the number of "subordinate" person entries in the authority document. These are structured person entries that, however, do not have a corresponding full-text description in the individual lexicons. As further described in Section 4, these are persons related to the main persons included in the lexicons, which were added manually to the authority document and linked to their main person. Finally, the table gives the sum of all the family and person records in the *SBI*, currently almost 9,000.

| | |
|---|---|
| *SBL* | 5,048 |
| *PSBL* | 1,839 |
| *NSBL* | 455 |
| *Σ* | 7,342 |
| *SBI-family* | 109 |
| *SBI-main* | 6,813 |
| *SBI-sub* | 2,032 |
| *Σ* | 8,954 |

Table 1: A quantitative overview of the SB database.

### 2.3 Structure of the authority records

The *SBI* authority document is encoded using the TEI module for names, dates, people and places (TEI, Chp. 13) that allows for their detailed annotation. Individual persons are encoded in <person> elements, while families are encoded in the <personGrp> elements.

[2] http://www.tei-c.org/

[3] https://desktop.github.com/

```
<person xml:id="sbi463215" corresp="sbl-text.xml#sbl02313" role="main">
   <idno type="URL">http://www.slovenska-biografija.si/oseba/sbi463215/</idno>
   <sex value="1"/>
   <persName>
      <forename>France</forename>
      <surname>Prešeren</surname>
   </persName>
   <persName>
      <forename xml:lang="de">Franz</forename>
      <surname>Prešern</surname>
   </persName>
   <occupation scheme="#occupation" code="#pesnik"/>
   <occupation scheme="#occupation" code="#pravnik"/>
   <birth>
      <date when="1800-12-03">3. dec. 1800</date>
      <placeName>
         <settlement>Vrba</settlement>
         <region type="municipal">Žirovnica</region>
         <country>Slovenija</country>
         <geo>46.3889184 14.146526</geo>
      </placeName>
   </birth>
   <death>
      <date when="1849-02-08">8. febr. 1849</date>
      <placeName>
         <settlement>Kranj</settlement>
         <country>Slovenija</country>
         <geo>46.2435206 14.3570883</geo>
      </placeName>
   </death>
</person>
```

Figure 2: TEI elements used in a typical authority record.

The structure of a typical person entry is illustrated in Figure 2. The <person> element is given its canonical ID and contains <sex>, <persName>, <occupation>, <birth>, <death>, further containing <date> and <placeName>. Dates have their ISO 8601 value in the @when attribute, while other TEI attributes are used in cases when the exact date is not known, e.g. <date notAfter="0940" source="#nsbl">pred letom 940</date> (*before year 940*, also giving the source of this information, in this case *NSBL*).

For the <settlement> element, we use gazetteers to ensure the standard form of settlement names. We also encode information about historical settlements for those that are not in existence any more or for settlements that still exist but have changed their name.

Of course, the actual structure, described above in principle, varies to a certain extent and depends on the information on a particular person; a detailed overview of all the elements used is given in Section 2.5.

The basic structured data was semi-automatically extracted from the corrected full text. First, regular expressions were written in Perl to extract pieces of information from the full texts and produce the initial authority records. For the most part, these regular expressions were quite simple, e.g. transforming "r. 3. dec. 1800" to <birth><date when="1800-12-03">3. dec. 1800</date></birth>. The automatically produced authority records were then manually checked and, where necessary, corrected.

It should be noted that even where no automatic annotation was possible, obligatory distinctions were tagged manually, such as different variants of names that were mentioned somewhere in the text of the article, important activities apart from the primary occupation of the person, or certain periods of time that marked important milestones in their life. The major aspects of this conversion process have been reported in more detail in Vide Ogrin and Erjavec (2007).

Where we were able to obtain high-quality images of the person, these are also included. More recently, we started to geocode (assign GPS coordinates) all place names. This is a four-step process that includes:

1. extracting the place names from the source XML document,
2. geocoding the extracted place names using Google's geo-location service[4],
3. checking the results for correctness and resolving cases returned no or multiple results, and
4. populating the source XML with the GPS coordinates.

To support the review process we developed a separate web application that allows the editors to quickly verify and amend the resolved location on an interactive map.

## 2.4 Occupation taxonomy

Along with the detailed markup, we also recognized the need for taxonomy of occupations. Its construction proceeded in a bottom-up fashion, i.e. we extracted and normalised the information from the occupations and activities that occur in the full text of the articles. The

[4] https://developers.google.com/maps/documentation/geolocation

taxonomy is already quite detailed with 1,077 categories, of which 160 have subordinate categories (e.g. “poklici-za-osebne-storitve”, *occupations-for-personal--services*), while 917 are leaf nodes (e.g. “brivec”, *barber*); the maximum depth of the taxonomy is 3.
Each category has, apart from its formal identifier, one or (in about 20% of the cases) two Slovenian glosses, giving also the name of the occupation for the female gender, as these often differ in Slovenian, e.g. “politik/političarka”, *male/female politician*.
The taxonomy is still work-in-progress, the most obvious reason being the fact that new persons are being added with the new *NSBL* volumes, so their occupations may be new to the existing taxonomy.

| Element | n |
|---|---|
| listPerson | 1 |
| personGrp | 109 |
| person | 8,845 |
| sex | 8,845 |
| idno | 6,921 |
| persName | 15,613 |
| forename | 12,379 |
| surname | 12,358 |
| name | 3,250 |
| genName | 78 |
| nameLink | 82 |
| birth | 5,817 |
| death | 5,371 |
| date | 11,414 |
| occupation | 13,107 |
| floruit | 224 |
| trait | 374 |
| roleName | 747 |
| placeName | 11,168 |
| settlement | 13,277 |
| country | 11,096 |
| geo | 8,948 |
| geogName | 36 |
| region | 2,584 |
| district | 597 |
| figure | 1,046 |
| figDesc | 1,046 |
| graphic | 1,046 |
| listRelation | 1 |
| relation | 1,735 |
| note | 750 |
| **Σ** | 158,865 |

Table 2: A quantitative overview of the TEI elements used in the *SBI* authority document.

### 2.5 Use of TEI elements

To give a comprehensive picture of the types and numbers of distinctions made, we list in Table 2 all the elements used in the body of the authority *SBI* document.
The elements are grouped roughly according to their function. The first group repeats the *SBI* information from the second part of Table 1, while the second group gives the basic information about a person, including the fine-grained classification of the parts of their names and information about their birth, death, occupation(s) and activities. The next group includes elements to do with locations, including the <geo> element giving their geocoding. Next comes a group that specifies the picture of the person, where available (currently just over a thousand). The final group contains the list of relations and the number of relations between persons that it includes, followed by 750 notes, giving some basic free-text information about a person, independent of the lexicon documents. All together, the authority document thus currently uses almost 160,000 TEI elements.

## 3. Implementation of the portal

### 3.1 Technologies used

Following the “XML as a data model” approach, we chose BaseX[5], a high performance and mature XML database server that provides rich querying mechanism through the use of XQuery[6] and XPath, as well as full text search via the integrated Lucene search engine. With this approach, we are able to use the original XML data directly in the database, with only minor structural transformations for performance optimization.
The implementation of the web portal and administrative interfaces is based on the Django[7] web framework, which in this case serves primarily as the “glue” between BaseX and the web interface. We also use Django to implement specific editorial tools like geolocation of place names.
The Slovenian Biography recently started to serve part of the data as JSON-LD[8] segments, which are embedded in the served HTML pages. JSON-LD is a new serialization format targeting Linked Data or Semantic Web Data that makes semantic annotation and publishing of data relatively easy. However, due to the lack of extensive biographical ontologies, it is not possible at this time to export a large part of the data that is otherwise available in the HTML (non-semantic) version. This lack of available specialized ontologies or schemas is also a problem when it comes to annotating and exposing the occupational taxonomy discussed in Section 2.4.

### 3.2 The Web front-end

The highly structured biographical data and metadata information is presented via user-friendly web pages of the portal. The TEI elements are rendered, which can be

[5] http://basex.org/
[6] https://www.w3.org/XML/Query/
[7] https://www.djangoproject.com/
[8] https://json-ld.org/spec/latest/json-ld/

rather complex for the detailed person names, which can use up to five different elements, and sometimes have several variants.

As shown in Figure 3, the available metadata is used for a number of aggregation and navigation views such as an alphabetical index, chronological index, browsable occupational taxonomy and interactive map, which all help users to explore the available data. The data is also searchable through the use of simple and advanced searches.


Figure 3: Navigation options of the SB portal: Search; Name index, Occupations and activities; Born / died on today's date; Families; Map.

## 4. Current work

The SB is a continuous work in progress, not only with new entries being added on a regular basis, but also striving to make the biographical information more access-friendly both to the average and to the more demanding, research oriented user.

We are also working on the enrichment of the encoding, in several directions. First, we are manually adding relations between persons, mostly family relations, but also close contemporaries, co-workers etc. This involves adding <relation> elements, which give the type of relation and references to the IDs of the related persons, e.g. <relation name = "parents" passive = "#sbi215416" active = "#sbi215416-0 #sbi215416-1"/>. As already mentioned in Section 2.2, we are also adding and linking new "subordinate" persons, which are related to the existing ones, to the authority document.

Second, named entities (i.e. person, location, organisation and "other") appearing in the text have been automatically annotated using the Stanford Named Entity Recognizer (NER) trained for Slovenian (Ljubešić et al., 2013). On a general and manually marked-up test corpus this NER tool achieved an overall precision of 73% and recall of 67%, however, the accuracy very much depends on the type of named entity: it is highest for persons (P = 82.2% , R = 86.7%) and lowest for the type "other" (P = 29.2%, R = 15.4%). This automatically assigned NER markup is now being manually verified and corrected.

Third, as can be already seen in Figure 1, the SB contains many abbreviations, which were numerous and typical in the print editions and then carried over into the digitised version, where they are now obviously unnecessary. We plan to do semi-automatic expansion of these abbreviations, where we are faced with two problems. First, the citation form of the abbreviation must be known, and, second, it needs to be inserted into the text in the correct inflected form, which is, of course, dependent on the context; as Slovenian is a highly inflected language, this is a difficult problem. We plan to approach it in the same way as the others, i.e. first using an automatic method to pre-annotate the abbreviations and then manually verify the results. We have currently manually annotated a sample of the SB containing 50 entries, and we will use this dataset to train a machine learning system to automatically expand and inflect the abbreviations – it should be noted that we also have the background lexicon containing most of the use abbreviations and providing their expansions to their citation form.

We are, as already mentioned, further elaborating our occupational taxonomy. The taxonomy IDS and the category descriptions are currently only in Slovenian, and, apart from adding new categories, further work will concentrate on translating the taxonomy to English and harmonising it with standard occupational taxonomies, such as SOC[9] (Standard Occupational Classification), a necessary step in the light of exposing our data also as RDF linked open data.

To improve search precision and provide a better user experience we are moving the implementation of the search to the Elasticsearch [10] that will allow us to

[9] https://www.bls.gov/soc/

[10] https://www.elastic.co/

implement more fine-grained and weighted full-text search, inclusion of a lemmatiser and a search type-ahead (autosuggest) search box.

## 5. Conclusions

The paper presented the Slovenian Biography, i.e. the source biographical lexicons that it contains, the process of their up-translation to the TEI encoded digital edition, methods used in editing and enhancing the data, the architecture and functionality of the web portal and the on-going work.

The SB is already extensively used: the analysis of access logs shows that in 2017 the portal had over 135,000 different users, who, during their 199,000 visits, accessed 332,000 pages, indicating that the SB is already perceived as a valuable and useful resource. Nevertheless, in our further work we also aim to focus on outreach, and on enriching its data, by linking the SB documents better with other Internet resources. We would like to connect the person descriptions of SB to other on-line biographical lexica, but also with relevant books, articles, pictures and multimedia content found on stable locations. We are also considering publishing the authority data in the scope of the CLARIN.SI repository under an open licence, so others can make use of it directly, thus enabling e.g. statistical (Anderson, 2007) or GIS-based investigations (Knowles & Hillier, 2008) over the data.

Finally, most users will likely search for person names via Google, and will typically first find the Wikipedia article of the person, where it exits. We have already taken the step of adding the external link to some SB articles from Wikipedia, and this practice should be continued and intensified, also adding stub articles to Wikipedia for missing persons and the link to SB to them. In this way, we also facilitate others to write the relevant Wikipedia articles.

### Acknowledgements

The authors would like to thank the three anonymous reviewers for their helpful comments and suggestions, which we have taken into account to the limit of our abilities. The work presented here was partially supported by the Slovenian research infrastructure CLARIN.SI and the Slovenian Research Agency programme “Knowledge Technologies”.

## 6. References

Anderson, M. (2007). Quantitative history. In W. Outwaite & S. Turner (Eds.), *The Sage Handbook of Social Science Methodology*. London: Sage Publications.

Cankar, I. et al. (eds.) (1925-1991). *Slovenski biografski leksikon*. Ljubljana: SAZU.

Erjavec, T., Ogrin, M. (2005). 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, 151-156.

Javoršek, J. J., Erjavec, T., Vide Ogrin, P. (2009a). Slovenian Biographical Lexicon – From a Digital Edition to an On-Line Application. In: *The Future of Information Sciences: Digital Information and Heritage: Proceedings of the 1st International Conference The Future of Information Sciences - INFuture 2009*. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu. 115-124.

Javoršek, J. J., Erjavec, T., Vide Ogrin, P. (2009b). The digitisation and deployment of the Slovenian Biographical Lexicon. *Research infrastructure for digital lexicography: proceedings of the 12th International Multiconference Information Society 2009*, Mondilex Fifth Open Workshop, Ljubljana, Slovenia, October 14-15, 2009. Ljubljana: Institut Jožef Stefan. 2009, pp. 64-71

Jevnikar, Martin et al. (eds.) (1974-1994). *Primorski slovenski biografski leksikon*. Gorica: Goriška Mohorjeva družba.

Knowles, A. K., and Hillier, A. (2008). *Placing history: how maps, spatial data, and GIS are changing historical scholarship*. ESRI, Inc.

Ljubešić, N., Stupar, M., Jurić, T., Agić, Ž. (2013). Combining available datasets for building named entity recognition models of Croatian and Slovene. *Jezikovne tehnologije*, (Slovenščina 2.0, ISSN 2335-2736, Tematska številka, Letn. 1, št. 2). Ljubljana: Trojina, zavod za uporabno slovenistiko. 2013, letn. 1, št. 2, pp. 35-57. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_03.pdf.

Svetina, B., et al. (2013-). *Novi Slovenski biografski leksikon*. Ljubljana: Založba ZRC, 2013-<2017>

TEI Consortium, eds. *Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/P5/.

Vide Ogrin, P., Erjavec, T. (2007). 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*. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu. 115-124.