**Преглед НЦД 24 (2014), 95–100**

**Jurij Hadalin**
Inštitut za novejšo zgodovino/
Institute of Contemporary History
Ljubljana

# A GLANCE AT THE DEVELOPMENT OF THE HISTORY OF SLOVENIA – SISTORY PORTAL AND THE ESTABLISHMENT OF THE WEB HUB SLOVENIAN DIGITAL INFRASTRUCTURE FOR ARTS AND HUMANITIES – SIDIH WEB HUB

**Abstract**: The following article provides insight into the activities of the Slovenian research infrastructures in the field of digital humanities. It describes the projects developed in the context of the Research Infrastructure of the Institute of Contemporary History and presents the ongoing efforts to establish the national infrastructure for digital humanities.

**Keywords**: digital humanities, Slovenia, report

## Introduction

The History of Slovenia – SIstory (www.sistory.si) internet portal reflects the work carried out by the Research Infrastructure of the Institute of Contemporary History in Ljubljana. The beginnings of the portal in 2007/2008 were relatively humble, and a modest digital library containing the collections of historical official gazettes and digitised issues of the most influential Slovenian scientific historical magazines (Zgodovinski časopis/Historical Review and Prispevki za novejšo zgodovino/Contributions to Contemporary History) represented the backbone of the portal. The main attraction for the wider public was a specific database containing the verified names of the victims of World War II and its aftermath in the Slovenian territory. These contents illustrate the initial purpose of the portal, primarily intended to cater to the needs of the Institute's programme and project groups. However, that stage was quickly exceeded. Today the portal provides access to a relatively extensive digital library, containing most of the Slovenian periodicals in the field of history and many works in the field of historiography, which is also interesting for the wider South-Eastern region. The publication of detailed databases and digitised sources – archival and printed – is even more valuable in terms of historiographical research. The rapidly-increasing internet connection speeds have also enabled us to publish a variety of video files, thus archiving numerous events in the field of historiography. As far as professionals are concerned, the History Citation Index database, which has now been in use for more than 6 years, is particularly useful. Since 2010 this database has been prepared in cooperation with the Archives of the Republic of Slovenia and Institute for Ethnic Studies. The History Citation Index contains a broad range of metadata and is an excellent vantage point for future metadata and content linking and upgrading, since it contains all the recent Slovenian works in the field of

historiography. At the next stage we are planning to enable the exportation of XML files and allow for the automatic creation of metadata whenever a new work is published on the portal.

## Basic Problems And Some Solutions Used

At the first stage much work and resources were invested in digitisation, and not so much attention was paid to the presentation. However, the rapid advances in the field of digital humanities stimulated us and helped us search for new preservation/presentation methods, as preservation is not our only goal. We are not merely interested in archiving, but also in the use/reuse of data. This process should be as painless as possible, but at the same time it should also facilitate complex research with the help of constantly-developed tools.

While attempting to ensure this, we had to tackle a complex and serious issue. All online materials presented on the portal are offered under the open-access policy, which brought the project into the realm of copyrights. Even a few years ago, during the period of frenetic digitisation, this issue was temporarily set aside. Furthermore, the legal framework of copyrights could also not envisage the rapidly-growing importance of the digital sphere. Under the current legislation the copyrights (or at least the material part of copyrights) remain in the domain of the authors until their death, after which they are passed on to their heirs for the period of as long as 70 years. This provision is not going to change anytime soon, especially if we take into account the most recent cosmetic changes of the copyright legislation at the European level. This brought an end to the previous frenzy, and some of our endeavours ended up in a dead alley. We soon started collecting written permissions from authors and their respective heirs concerning the republishing of their works in the digital format. This was only possible due to the very straightforward and concise form and the fact that Slovenia is so small. Considering the size of the Slovenian historiographical community, we can usually get in touch with the majority of authors. Attempting to establish contacts and ensuring positive responses can be a painstaking but also rewarding endeavour, since we can also offer the newer works to the public. Meanwhile, the other alternative would leave us without an extensive collection of 20$^{th}$-century copyrighted interpretations.

The portal is based on an actively-administrated MySQL database, and new publications and metadata are being entered into it on a daily basis. The metadata was one of our main concerns in the recent years, since this issue came into the foreground after a short delay. If the main preoccupation at the first stage of the development was to digitise, metadata was an integral part of the second stage. It was presented on the new portal, which has been in operation since October 2011. The new version of the portal was also a chance to address the problem of URN, which is essential but was not considered necessary at the beginning. Some of what are today obvious elements and solutions have been integrated into the portal during the course of its development, following the rapidly-expanding and hard-to-organise quantities of content. Perhaps this sequence is somewhat unorthodox when planning a portal, but a critical mass of data is necessary to provide the creators with an in-depth knowledge of the actual requirements, thus helping to find a suitable solution. The vantage point was aimed low, with the Dublin Core Metadata Standard as key, adding some original fields as necessary. At the time of writing this scheme is already much broader: the original DC standard contains 15 mandatory fields, while with the last upgrade 50 fields were exceeded. This expansion is a result of our cooperation with different organisations in the field of history, all of which have their own specific demands. A very simple example that occurred during this endeavour is the archival signature.

Fortunately two years ago the quantity of the published material was still manageable. The new database provided a fairly simple way to divide copies of different periodicals and other publications containing the works of multiple authors into smaller units using the **is part/has parts** options in the DC. The originality of the volumes thus remained undisturbed and allowed for a very popular option among historians, who are used to leafing through periodicals.

Another potential which can easily be overlooked is hiding in the use of metadata. Already in the beginning the SIstory portal was constructed as bilingual (Slovenian/English). The Slovenian pages provide a wider choice, but we also tend to input the English metadata whenever possible. Bilingual metadata ensures a better visibility and search ability of the portal's content in the web browsers as well as far better options of placing our metadata into the pan-European networks. The success is apparent, as the number of visitors doubled from 2011 to 2012 and is nearing 60.000 unique visitors annually.

It took one person more than a year to manually enter the required metadata, which then also allowed for the development of a functional internal search with corresponding filters, combining metadata with full-text search. The OCR of PDF files, which has gradually improved, then led to the idea of upgrading our metadata model simultaneously with the Text Encoding Initiative metadata model. In the first half of 2013, as advised by the experts from the Josef Stefan Research Institute, most text files will be transformed into TEI files via XSLT (mostly level 1 in accordance with the TEI recommendations for libraries, but some also on higher levels). The DC standard metadata will be transferred into the header of the TEI file, while the text in the XML file, extracted with OCR, will constitute the core of the document. This method should provide a clear insight into the quality of OCR, which still had much room for improvement during the first years of digitisation. Some of the PDF files will be removed and upgraded in the process, but the most important advantage of this effort is a desire for the long-term preservation of text, not only images, as it has been customary until now. One of the highly desirable side effects is also the ability to provide the instant ("on-the-fly") conversion of text into formats like ePub or Mobi. The availability of different text formats will ensure easy access for the important segment of smartphone and tablet users.

## Offspring (Sub)Projects

The emerging share of smartphone/tablet users was detected as a particularly important target group, which is why a pilot version of the mobile portal was developed in 2011 (www.sistory.eu). This „beta“ version provided user-friendly browsing of monographs and serial publications in suitable formats. The open-source tool jQuery mobile (http://jquerymobile.com/) served as a platform. At the end of the same year another „beta“ project, called ZGOLj (Zgodovina Ljubljane/History of Ljubljana), received a favourable response from the wider public. This application is based on the Layar augmented reality browser, which can be used on Android or Apple OS smartphones, offering the users the option of embarking on sightseeing tours of Ljubljana. Old photographs, pictures and architectural plans originating from the Ljubljana Historical Archive, stored in the SIstory database, were reused in a slightly different manner. After the users download the application, they are able to view photos and plans for every building in the Ljubljana historical city centre with the help of their integrated camera (geolocation represents a part of the SIstory metadata). The graphic material is accompanied with descriptions of the historical importance of buildings, even those which no longer exist.

The wide network we have established with our partners allows for widespread digitisation as well as publication of original digital resources. However, as described in the previous paragraph, we tend to present the materials in various ways. This is made possible by the flexible architecture of the portal. One of the recent examples of such practice is the presentation of the Stenographic Minutes of the National Assembly and the Senate of the Kingdom of Serbs, Croats and Slovenes/Yugoslavia, originating from the archival fund kept in the Archives of the Republic of Slovenia. After the material was digitised, we have released all of the publications under the label of archival sources as they were originally bound and in the same order as they were kept in the archival boxes. Using the principle of breaking down larger units into smaller ones using DC metadata, which we have already mentioned, we have gained access to separate units/sessions. This enabled us to present the content according to archival standards and link it to the archival SCOPE network, which is – unlike other systems – organised hierarchically. As such arrangement does not allow for the research logic which historians are accustomed to (a good example of this logic can be found at http://alex.onb.ac.at/), we came up with a new design, enabling researchers to browse through parliamentary sessions by date.

A similar selection of open-source tools was also used in order to prepare a presentation of the census of the Jewish population in the territory of the Drava Banate in 1937. This relatively small population census was presented on a platform based on the jQuery tool (http://jquery.com/) and HTML5. In the course of that endeavour, short document transcriptions were built on the basis of the OMEKA tool collection (http://omeka.org/).

The most ambitious application of the ICH RI is a longitudinal project of population censuses. The first preserved census of the population of Ljubljana from 1830 has been available on the SIstory portal since 2010. The database combines transcription and scanned census sheet (originally in the German Gothic alphabet), enabling proofreading and allowing the users to extract additional information. In the beginning of 2013 the database should be significantly upgraded and enriched with a new series of population censuses, namely those carried out in Ljubljana in 1857 and 1869, as well as partly with the last census before World War II, carried out in 1931. This database is not limited only to Ljubljana – the population census forms from Novo Mesto and Izola have already been digitised as well, and they are currently being transcribed. In the future the project is intended to cover the entire Slovenian territory. Population censuses represent a valuable historical source, not only for demographic studies, but also for genealogical research. All of the individual handwritten forms have been preserved, which is a rare occurrence in the Central and Eastern Europe as all of the original material was usually destroyed after it had been used in the statistical analysis at the time in order to anonymise the information. The forms contain extensive information about social status, vocations, hygiene (descriptions of toilets, water sources), and even list domestic animals. Each head of the family filled this form for all other family members, and we can trace these people throughout the 19$^{th}$ and 20$^{th}$ century. Due to the protection of personal data (such information has to be protected for 75 years), the project will conclude with the 1931 census.

Most of the material was scanned on microfilm 20 years ago, since it was one of the most sought-after collections at the archives. New scans of the original documents were made (colour, 400 dpi). The process of transcription/transliteration has been in progress for a year now, since it involves a lot of "manual" work. The obsolete German Gothic alphabet and the use of German in the documents until the 1880's presents a special issue. Every form represents a household, which is transcribed into a metadata scheme. Each scheme can be exported as an XLS file. The database also makes it possible to read the attached scanned

original which is an integral part of the database, and a magnifying tool is provided as well. The quantity of data to be transcribed is enormous, and only a part of this work can be completed within a reasonable time frame by the colleagues at the Institute. Due to this issue we have come up with the idea of crowdsourcing by enlisting the participation of the students of history at the Ljubljana University. The ultimate goal is to link all that data with other administrative archival collections, which would enable a complete analysis of individual family trees.

## Emergence of the Sidih Portal

The achievements in the field of historiography and the established network of partner institutions allowed us to make a further step in the popularisation of DH: the construction of a portal called Slovenian Digital Infrastructure for Arts and Humanities – SIDIH (www.sidih.si). ICH – in partnership with the Scientific Research Centre of the Slovenian Academy of Science and Arts – is the Slovenian partner of the European network for digital humanities DARIAH (www.dariah.eu). The membership in this pan-European network enabled us to identify the critical problems of the Slovenian DH. Despite a rapid advancement, the fundamental problems of creating and preserving metadata as well as copyright management remain unsolved. Digital humanities have been established in Slovenia for a long time, but most of the results of these activities remain scattered throughout different institutions – some of them are accessible to the public, while others are only used internally or even kept on personal desktops. The extensive collection of original digital sources has been enriched with a vast quantity of digitised sources (especially at the institutions under the auspices of the Ministry of Culture). In many cases the digitised material has only been stored and has not been prepared for publication.

The process of preparing high-quality DH contents calls for regular maintenance and creation of appropriate metadata. This is the only way to keep perfecting the search engines, ensure the functioning of new tools, and allow researchers to benefit fully from the digital contents. This was the basic idea behind the SIDIH portal, which represents a national entry point to DARIAH and also includes a browser of the Slovenian DH material. The browser should be an incentive for all the institutions to prepare their contents in a suitable manner. The architecture of the portal is fairly uncomplicated and mostly based on open-source tools. A harvester using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) represents the backbone of the browser. Considering the different levels and variety of the available content, a simplified DC metadata standard is currently in use as the smallest common denominator. SIDIH only provides researchers with the basic information about the sources, and researchers are then able to find additional metadata and contents in special repositories. Already in the first months of the trial period (second half of 2012) it became evident that researchers sought information not only in the digitised materials, but also made use of the metadata about analogue sources. A need to define different types of materials and implement user rights has also emerged. Strict adherence to the DCMI Type and, in addition, to an optional controlled vocabulary of type (depending on the needs of particular disciplines) would solve this problem. The established needs for harvesting, processing and presenting the metadata dictate a limitation of DC elements. Therefore a special SIDIH metadata model (an upgrade of the DC XML scheme) should be implemented.

Thanks to the ongoing improvements and better metadata processing, SIDIH should enable a search engine upgrade and use of diverse tools. The foundations of the project at the initial stage are represented by the contents of the two largest national DH repositories, the

SIstory portal and @rzenal (SRC SASA - http://www.arzenal.si/). The network is going to be expanded in 2013 with the TEI based portals eZISS – Scholarly Digital Editions of Slovenian Literature (http://nl.ijs.si/e-zrc/), NRSS – Unknown 17th and 18th Century Manuscripts of Slovenian Literature (ezb.ijs.si/nrss/), and the Slovenian Biographical Lexicon (http://nl.ijs.si/fedora/get/sbl:sbl/VIEW/).

The main goal of the SIDIH portal is to become a hub for Slovenian digital humanities and to encourage researchers to use new technologies. A list of Slovenian DH projects, past and present, is in the making, and we also tend to publish Slovenian translations of some of the most relevant documents from the DH world as well as prepare collections of links to the best-practice projects and DH initiatives.

jurij.hadalin@inz.si