Digitization as a Means of Preservation of Russia's Literary Heritage - a case study: The Digitization of the Letopis' Zhurnal'nykh Statei

Presented at the "Managing the Digital Future of Libraries" conference, Russian State Library, Moscow Russia, April 17-19, 2000.

George Andrew Spencer
Visiting Assistant Librarian
Russian Periodical Index Project Manager
Digital Library Program
Indiana University

This paper was published in the "Russian Digital Libraries Journal, vol. 3, issue 4 (2000):
http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2000/part4/spencer

Russian translation available at:
http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2000/part4/spencer

---

Indiana University has a strong history of Slavic studies dating back more than half a century. A Russian Studies Department was founded in the autumn of 1947, which in 1958 became the Slavic Languages and Literatures Department and the Russian and East European Institute was founded also in 1958. As a result of this strong focus on Slavic studies, the Indiana University Library system has a strong Slavic collection.

The Indiana University Digital Library Program was founded in August 1997. Since its inception the Digital Library Program has been active in developing digital projects in music, photography and English language texts.[1] Thus with Indiana University having interests in Digital Library initiatives as well as Slavic studies, it was natural that the Indiana University should embark on a Russian language digitization project. In 1999 the Digital Library Program was awarded a three year United States Department of Education Title VI Technology Program grant to explore the issues associated with using digitization as a means of preserving embrittled high-acid paper Russian language library materials and making the resulting Cyrillic text freely available worldwide on the Internet.

The Letopis' zhurnal'nykh statei was chosen for this project in consultation with Slavic bibliographers throughout the United States who were asked to suggest a Soviet era serial that was most worthy of being preserved but that had not already been preserved by other means. The Letopis' zhurnal'nykh statei was particularly recommended for digital preservation because of the difficulty in using the print original. With only author name and geographical name

indexes in the print original, it was felt by a number of bibliographers that adding keyword searching via a SGML or XML search engine would significantly enhance both the usability and the usefulness of the Letopis' zhurnal'nykh statei to scholars around the world. The preservation of the Letopis' zhurnal'nykh statei is particularly important since many of the journals indexed within it have been preserved on microfilm, and thus will continue to be available to scholars for the foreseeable future. Therefore it is important to preserve this major journal index.

Once the Letopis' zhurnal'nykh statei was chosen as a candidate for digital preservation, in keeping with Indiana University's strong commitment to intellectual property rights, legal council was sought to determine the copyright status of the Letopis' zhurnal'nykh statei. It was determined that under both current Russian and American copyright laws that the Letopis' zhurnal'nykh statei was in the public domain.

It was decided that given the budget and time constraints of the grant a twenty-year span of the Letopis' zhurnal'nykh statei could be digitized and this span would be 1956-1975. This was based upon the fact that Kraus Reprint of Liechtenstein had in the period 1969-1970 reprinted the Letopis' zhurnal'nykh statei issues from 1926 through 1955. As a result, it is beginning with the 1956 issues that preservation of the Letopis' zhurnal'nykh statei is most needed.

As is well known, Soviet era documents were, for the most part, printed on high-acid paper. Indeed, judging by the issues in Indiana University's collection, acidic paper continued to be used for the Letopis' zhurnal'nykh statei as recently as mid-1999.

Traditionally there have been only a few strategies for the preservation of library materials that have been produced on high-acid paper. Deacidification and conversion to microform formats were the two main options normally under consideration until recently. However, although some deacidification processes claim limited strengthening of brittle paper, paper that has already become very brittle is usually not considered for deacidification.[2] Thus until recently transferal of the intellectual content to microform formats has been the method of choice for preservation by the library community. This in spite of the fact that the microfilm format in particular is inherently difficult to use, with lack of random access and eye strain among the difficulties associated with the format.[3] Now as the field of Librarianship transitions to a more digital environment, the possibility of digital preservation of the intellectual content of high-acid brittle paper originals is being actively explored. For an index, such as the Letopis' zhurnal'nykh statei, the lack of random access inherent in microfilm is particularly a problem for efficient usage. For resources such as the Letopis' zhurnal'nykh statei, digitization seems a more appropriate option as a method of preservation than microfilm.

At the digitization phase of the project, decisions had to be made as to the color depth of the images to be made. Scanning the originals as grayscale images allows the yellowed paper to become apparent in the background of the image. This is a disadvantage since this can have an impact on the accuracy level of the Optical Character Recognition (OCR) software used to convert the digital images back into text files. It was found that with extremely yellowed paper the background paper is dark enough to be sometimes interpreted by the OCR software as spurious faint text. On the other hand, scanning as bi-tonal images, which could perhaps result

in the occasional loss of very faint characters, still seems to result in better overall OCR accuracy because of the removal of the yellowed paper background noise.

Overall image file size is also related to choice of color depth. A grayscale image of the same original scanned at the same resolution will obviously result in a larger file size than a bi-tonal image. In addition to color depth, resolution is the other factor with an effect on file size. In the U.S., resolution is usually measured in terms of "dots per inch" (d.p.i.). Tests were run on samples of Cyrillic text scanned at 300 d.p.i. and 600 d.p.i., two of the standard resolution settings available on most scanners, which showed only a marginal difference in the accuracy rate of the OCR software when using 600 d.p.i. versus 300 d.p.i.

It was found that a page of the Letopis' zhurnal'nykh statei scanned bi-tonally at 300 d.p.i. yielded an image file of .48 megabytes while the same page scanned as 12-bit grayscale at 600 d.p.i. yielded an image file of 14.49 megabytes. In addition, since we chose the Tagged Image File Format (TIFF) for our image files, we also specified that the files use ITU (formerly CCITT) type 4 compression to further decrease the file size.

If digital text files are the only result needed, 300 d.p.i. would perhaps be adequate and would have the advantage of smaller image file size. In addition the time per image to scan an original is also dependent on the resolution required. Hence in large projects such as this scanning at higher resolution can significantly increase the labor costs for scanning. Also if outsourcing of the scanning is to be considered, for some vendors the cost of digitization is dependent on the resolution required. Thus, 300 d.p.i. could also be cheaper on a per page basis as well.

However in the case of our project, it was decided that since these digital images may someday be used to also produce new paper copies, it was better to use the higher resolution of 600 d.p.i. in order to have higher quality images from which to produce paper copies directly. In spite of using the higher resolution, because we are using ITU type 4 file compression in conjunction with bi-tonal color depth, we are achieving file sizes for a page of text on the order of 100 kilobytes.

It was decided that given the challenges of the OCR and encoding segments of the project that the actual scanning of the originals would be outsourced, if possible. Cognizant of the success of other digital library projects which have outsourced the digitization phase of their projects, it was decided that this was the most efficient use of our resources.[4]

When dealing with over 200,000 images a standard file naming convention was essential to maintain control of the inventory of the TIFF images and to monitor the workflow. We decided in order to ensure long-term access to the data to follow strict ISO-9660 standard file naming conventions. Within this constraint, we specified that the file name for each file (that is, each page image) would contain the year, issue number and page number. In addition, we specified several data fields within the TIFF file header to be populated with data chosen to ensure long-term inventory control over the image files as well as other technical data on the scanner settings and file compression used to create the image and archival data such as the date of scanning.

The originals were disbound by the Indiana University Library Preservation Department and the binder's edge trimmed before shipment to the scanning vendor. The scanning vendor scanned the originals, and placed the resulting TIFF images on CD-R media. In addition to the CD-R media, the originals were also returned to us, to be used in the proofreading stage of the OCR process.

Once the TIFF image files were in hand, the process of Optical Character Recognition could begin. At this stage a decision needed to be made as to which character encoding in which to save the OCR output text files. It is well known there are several competing encodings for Cyrillic text such as KOI-8, Microsoft CP-1251, and Unicode [UTF-8] among others. While there do exist utilities [example: uniconv] that can convert files from one code page to another, when dealing with many thousands of pages it is obviously best to produce the pages in one encoding and stay with this encoding throughout the process.

This decision also has an impact on the choice of SGML/XML search engine to be used for the project. Since currently the majority of Indiana University computer workstations use Windows NT 4.0, it was decided that Unicode would be our best solution. Unfortunately, a problem with Unicode is that although it is part of the XML specification it is not currently part of the SGML specification.

It was found in early testing that the SGML search engine that had been used for earlier Indiana University Digital Library projects did not work well for Cyrillic materials. It was eventually decided that if a XML search engine could be found that was scalable to this size database and available at a reasonable cost given the budget constraints of the grant, that the built-in Unicode compatibility of XML would facilitate the markup stage of the project.

Another consideration impacting the choice of a search engine is the use of a commercial search engine versus an open source search engine. If an open source search engine is used, then the option of later distributing the database and its search engine in CD-ROM format is maintained. Distribution of the database on CD-ROM is potentially a more useful format to libraries with limited band-width Internet connections. On the other hand if a commercial search engine is used then licensing and associated costs could possibly prohibit any CD-ROM distribution.

Once the encoding of the OCR output text files had been chosen then the OCR phase of the project could begin. It has been found that there are several considerations that affect the accuracy rate of the OCR software. One problem is that the wicking effect of the ink into the fibers of the paper causes certain punctuation to be indistinct. Specifically, periods are often mistaken for commas and colons are mistaken for semi-colons by the OCR software.

A second problem is recognition of non-Cyrillic non-Roman characters. This is particularly a problem in the exact sciences where Greek characters are commonly used for variables and the Roman alphabet is used for Latin species names. The OCR software we are using has a Russian/English mode; however it was found that the accuracy rate of the Russian text decreased slightly when the OCR software was used in the Russian/English mode instead of the purely Russian mode. But overall, especially in the exact sciences sections of the Letopis'

zhurnal'nykh statei, the benefit of running the software in the Russian/English mode in order to recognize the Roman alphabet strings outweighed the cost of slightly lower accuracy rate for the Russian text. But the OCR software we are using has no mode which will recognize Greek characters as well. Another analogous problem is the recognition of special characters and sub-scripts and superscripts which are particularly common in the Mathematics and Chemistry sections. As a result of this it was found that the exact sciences sections of the Letopis' zhurnal'nykh statei required far more intensive proofreading and correcting than other sections.

Meeting the challenge of marking up more than 200,000 pages of text is aided by the use of Perl programs which use regular expressions to match and substitute such fields as the journal title. A database was compiled of all the journal titles present in the Letopis' zhurnal'nykh statei for the years being digitized, and a Perl program is ran which matches the journal titles in the database against text strings in the OCR output text files and replaces them with SGML/XML tagged strings.

Unfortunately, the problem with the OCR software mistaking periods for commas also has an impact on the regular expression Perl program for the SGML/XML tagging. For example, the Perl program is expecting a period in the character string being matched and not finding a match because the OCR software has rendered the character as a comma.

Perl programs are also used for such tasks as concatenating the individual OCR output text files into larger files which are more easily handled by the search engine and removing extraneous hyphenation caused by word breaks at the end of lines in the original text. All through this project one of our goals has been to try to automate as much of the process as possible, to test the proposition that digitization can be a viable method for large scale preservation of Russian language library materials both in terms of time and money resources.

A final consideration in using digitization as a mode of long-term preservation of library materials is maintaining the integrity and accessibility of the data. In our case, once the project has been completed, the University Information Technology Services division of the University has undertaken to maintain the data for the long term and to perform data migration as new storage systems and storage media come into use. Thus this database has the potential to become a long-term preservation copy of the Letopis' zhurnal'nykh statei.

For more information on our project, also see our web site:
http://www.dlib.indiana.edu/collections/letopis/letopismain.html

Footnotes

1. Indiana University Digital library web-site: http://www.dlib.indiana.edu/main.html

2. Cloonan, Michele Valerie "Mass deacidification in the 1990s" in Harvey, Ross. **Preservation in libraries: a reader.** New York : Bowker Saur, 1993. pp. 376-378

3. Shoaf, Eric C. "Preservation and Digitization: Trends and Implications" *Advances in Librarianship* vol. 20 (1996) pp. 223-239

4. Cornell University / University of Michigan "Making of America project" (http://moa.cit.cornell.edu/MOA/moa-mission.html), and the "Historic Pittsburgh project" (http://digital.library.pitt.edu/pittsburgh/vendors.html)