PRACTICAL (AND IMPRACTICAL) ASPECTS OF BUILDING THE "MEETING OF FRONTIERS" RUSSIAN-AMERICAN COOPERATIVE WEB SITE WITH RUSSIAN PARTNERS

Michael Neubert, Reference Specialist, Library of Congress
Library of Congress, European Division, 101 Independence Avenue SE, Washington DC 20540-4830, mneu@loc.gov
Deborah M. Thomas, Digital Conversion Coordinator, Library of Congress
Library of Congress, National Digital Library Program, 101 Independence Avenue SE, Washington DC 20540-1310, deth@loc.gov

Cooperation of any kind between libraries presents challenges. This is a case study look at cooperation between the Library of Congress (LC) and several Russian institutions, including the Russian State Library (RSL), the National Library of Russia (NLR), and local institutions in Western Siberia (none of which are actually libraries) on creating the "Meeting of Frontiers" digital library web site. [http://frontiers.loc.gov/] The Library of Congress considers that the "Meeting of Frontiers" project is a success. In keeping with its stated goals, LC has acquired digitally-converted materials from Russian partners to build a thematic digital library. Much of the technical information and description of digitization practices is available on the Library of Congress American Memory server - see http://memory.loc.gov/ammem/ftpfiles.html and http://frontiers.loc.gov/intldl/mtfhtml/mfabout/mfdigitz.html. The focus here, however, is on a number of more problematic areas: interoperability issues between institutions, implementation of "standards," and selection of materials for digitization.

Project Background

"Meeting of Frontiers" is a U.S. Congress-funded "bilingual, multimedia English-Russian digital library that tells the story of the American exploration and settlement of the [American] West, the parallel exploration and settlement of Siberia and the Russian Far East, and the meeting of the Russian-American frontier in Alaska and the Pacific Northwest." [http://frontiers.loc.gov/intldl/mtfhtml/mfhome.html] The site is intended for use by school children and the general public as well as scholars. LC has digitized materials from its own collections on the relevant themes and is now working with partners, primarily in Russia, to digitize materials from those institutions to integrate into the Meeting of Frontiers site.

The following is a chronology of the developments with Russian partners:

- December 1999 – The Russian State Library and the National Library of Russia sign agreements to work with the Library of Congress on building the "Frontiers" web site. The agreements describe technical aspects of the project, equipment to be used, and how selection will take place.

- May 2000 – Equipment purchased by the Library of Congress is installed at the Russian State Library and the National Library of Russia. LC staff provide training to RSL and NLR staff who will perform scanning. Scanning (production) begins.
- January 2001 - First update of the "Meeting of Frontiers" site on the Library of Congress server using images and description from Russian partners.
- April 2001 - The Open Society Institute (OSI) signs an agreement with the Library of Congress to coordinate the participation of regional institutions in Western Siberia (Novosibirsk and neighboring oblasts) and coordinate digitization production at these sites.
- June 2001 - Delivery of equipment to Novosibirsk for the Novosibirsk Mobile Scanning Laboratory to be operated by OSI and funded by LC with LC equipment, to conduct scanning at regional institutions (museums and archives) in Western Siberia.

The following is an overview of the cooperative aspects of this project (who is responsible for what, who does what):

- Agreements (both LC and partner organizations)
- Equipment (purchased by LC, serviced by LC, with partner cooperation)
- Selection of what to digitize (both LC and partner organizations)
- Scanning (done to "standards" provided by the Library of Congress – training provided by LC staff to explain these "standards")
- Description (bibliographic description following institutional practice; formatting of the data in Microsoft Access according to LC "standards" – training provided by LC staff to explain them)
- Adding the materials into the existing web site (all the following tasks are the responsibility of the Library of Congress – note the extensive use of LC staff resources to accomplish them)
  - Quality review of images (and feedback to Russian institutions)
  - Creation of derivative files (JPEGs, GIFs from TIFF files supplied by partners)
  - Organization of files for retrieval on the LC server
  - Creation of original text (and incorporation into the existing site) – both in English and Russian
  - Drafting of HTML
  - Creation of indices linked to images and description
  - Development of interface program "features" (based on conceptualization of presentation)

Interoperability Issues Between Institutions

The Library of Congress and its partners have had to solve a number of technical problems in order to build the "Meeting of Frontiers" web site. One aspect of the technical approach used was the determination of the level of detail describing technical aspects of the project presented in the cooperative agreements, particularly with the Russian State Library and the National Library of

Russia. The agreements include substantial appendixes giving specifications for files to be created, for the format of the bibliographic description, and describing the technical characteristics of the equipment used, but these parameters are more of a statement of intentions than a strict set of rules to be followed blindly. The specifications are based on those used in all LC digital access projects.

The equipment purchased for use in this project is of sufficient quality to insure that the participating Russian institutions would be able to produce images of similar quality to those produced at the Library of Congress. It was also considered important that the technical solution used be one that Library of Congress staff were familiar with – thus it was decided to digitize the materials using an approach that centers around different models of PhaseOne digital camera backs (one model is used is both the Russian State Library and the National Library of Russia, another more portable unit will be used in Western Siberia). The Russian State Library and the National Library of Russia have, in addition to the PhaseOne equipment, a single Agfa T2000 flatbed scanner for scanning flat materials. The limited choice in equipment limits efficiency in scanning production (that is, if more types of scanning equipment were available, production would go faster in some cases) but the chosen equipment allows for digitizing most types of materials and what was possible given the limited budget for purchasing digitizing equipment.

Some of the technical issues that were overcome would seem trivial to Russian web designers, but in the context of the AIX server environment used at the Library of Congress to manage its web site, this was not the case. A particular example is the integration of Cyrillic text (using Windows Codepage 1251) into the static HTML. There have also been problems with transmission of the Cyrillic text within the Library as pages have been developed.

The materials from the Russian partners have had qualities that have effectively forced the Library of Congress to develop new types of presentations on the "Meeting of Frontiers" web site (presentations not used by the American Memory site). This has been true of several of the photograph albums contributed by the Russian State Library and the National Library of Russia. See http://frontiers.loc.gov/intldl/mtfhtml/mfdigcol/rslph.html or http://frontiers.loc.gov/intldl/mtfhtml/mfdigcol/nlrph.html for examples.

There are also some issues that are not yet solved. For example, while the static HTML is fully bilingual, it has not been possible to integrate Cyrillic text (other than transliteration according to LC standard) into the underlying database of individual items. In other words, when a user looks at an individual item on the site, the associated record that displays with the item contains no Cyrillic text, although there are often fields containing Russian language entries – in these records the Russian is presented in transliteration.

A related problem that has not been solved is the inability to search the web site using Cyrillic characters. This is true both of the static HTML, which does contain substantial quantities of Cyrillic text as well as within the descriptive records – although as noted, the records do not contain Cyrillic Russian but rather transliterated Russian. The complications of bilingual searching (stemming, word forms, stopwords, and so on) have prevented LC from developing a bilingual search tool for use on the AIX server.

A problem that may be difficult to solve that sits now on the horizon is the interest of many of the partners to establish various kinds of what might be most accurately called "pseudo-mirror" "Meeting of Frontiers" sites (not true mirror sites, but ones that would contain much of the content now on LC's server). There are many issues, but probably foremost is the difficulty in replicating the development of the LC web delivery application as a complete system for over seven million items and architecture specific to LC production workflow for outside use. Any mirror site will probably need to develop customized software to display and present digital materials of varying formats and specifications.

Standards

The Library of Congress's National Digital Library Program has been a leader in developing approaches to presenting its materials via the Internet through its American Memory web site. [http://memory.loc.gov/] Considerable amounts of technical information about how this has been done are presented on the site. One drawback of such leadership, however, has been that the Library of Congress has developed and implemented solutions that have not necessarily gone on to become standards (a "standard" here being understand to be a solution or approach that has been general approval to serve as a model). This is why we refer to LC practices concerning technical requirements for digitization and description as "standards" (that is, pseudo-standards) since they are not true ISO (or other) standards.

For "Meeting of Frontiers" the Library of Congress has felt it necessary to require some adherence by the Russian partners to LC's own existing "standards" – but this has not been strictly applied, in part because the relationships with the partners are supposed to be cooperative (rather than LC dictating terms) and in part for practical reasons. For digitization, for example, the partners have been allowed greater latitude in setting the resolution of files that they create than would be given to a (commercial) contractor producing the same files for the Library of Congress (in such a contract, resulting files are strictly governed by the terms of the agreement). This is largely because of practical problems given the limited (although high resolution) ability of the equipment they have to produce files as LC would otherwise specify.

The description of the materials that is presented on the site has two aspects: the "rules" (or standard) for description and the format for the data (metadata). Because this is a cooperative project, the Library of Congress did not wish to attempt to impose some arbitrary standard for description upon the Russian participants. Instead LC requested that certain minimal types of information be supplied and that this description can follow whatever rules of description that the institution would choose to present such information about its collections via the Internet. LC has requested that each institution should apply whatever rules consistently. (A result is that the Russian institutions follow the Russian practice of modernizing orthography for pre-revolutionary materials while LC's records use the orthography as it appears on the original piece – all in transliteration, of course. But this has practical implications for searching of the descriptive records – for example, русского versus русскаго, and so on.) At the same time LC has kept tight control of the format of the database

records (maintained at the working level as a Microsoft Access 2000 database). This is a practical necessity to allow integration of data received from multiple sources. The use of more common library description standards, such as MARC, Dublin-Core, or SML markup have been avoided because of inter-institutional practicalities; variations in actual use of MARC-21, RusMARC, and XML led to the decision to provide a simple fielded database for data that would provide simple interoperability with the existing database already web accessible at LC.

It is, of course, an open question as to whether this can be considered a "cooperative project" when the Library of Congress has so much control over the standards (or "standards") being used. The materials must be integrated into an existing web site approach – the Library of Congress has tried to minimize reworking long-established practices when presenting partners' materials on the "Meeting of Frontiers" web site. (Note that the Library of Congress has gone to great lengths to solicit and incorporate suggestions from Russian participants and specialists on the presentation of Russian history that is available from the site, including a May 2000 conference in Alaska that included many Russian participants.)

Selection

In its initial project development, LC outlined a thematic scheme for describing the history of the "Meeting of Frontiers" (that is, the history of the development of the subject regions in Russia and the United States). This scheme was intended to be modular and flexible and to provide ample possibility for additional and amendments as partners joined the project and contributed digital materials. In selecting materials, partners are asked to review their own collections and determine what may add to or complement materials already online. LC maintains final approval of these selections for several reasons: without central organization, duplication of digitized materials may result; some materials fit better into the themes better than others; some physical formats may prove more resource intensive or difficult to digitize than they are worth for the project; and some materials my be problematic to display or provide access to within the existing LC digital library.

An issue tangential to selection is the need for LC's partners to supply LC with certification that materials digitized are free of restrictions under Russian copyright law. The Library of Congress is the home of the United States Copyright Office and much weight is given to assuring compliance with all relevant laws. As a practical matter, this has meant that most of the materials digitized have been older ones, in the public domain, however in some cases partners have digitized modern materials, securing authorization in writing from the Russian copyright holder.

Educational Goals

"Meeting of Frontiers" is intended for use in schools as well as for use by the general public and scholars.. The Library of Congress's American Memory site has its companion "Learning Page" site (at http://memory.loc.gov/ammem/ndlpedu/index.html) but at present "Meeting of Frontiers does not have an analogous page to help educators use "Meeting of Frontiers" materials. The educational

goals of the project are unlikely to be met through such LC-produced materials at the present time. Differing educational standards and traditions mean that intensive efforts (beyond what the institution is capable of) would be required for LC to produce these materials on its own. In general the site is aimed at the broadest possible audience, providing illustrative maps and timelines for use in schools – but without providing lesson plans, discussions questions, or activities. LC hopes that partnerships that may be established with outside organizations (and the additional resources that these partnerships may bring) will allow the development of such materials in the future.

Analogous Operations

The "Meeting of Frontiers" digitization effort is not the only cooperative project by which the Library of Congress is acquiring copies of materials in Russian libraries – LC also has established a cooperative microfilming program with microfilming equipment in a number of institutions, including the Russian State Library. For the most part the microfilming operations have been far less complex than digitizing of materials for "Meeting of Frontiers", primarily because they much more closely resemble traditional library acquisitions. The partner institutions where microfilming is performed offer lists of materials and LC makes selections for what will be microfilmed - not according to the needs of a web site but using more traditional library selection criteria. (The question of what happens at LC with the digital files that are being "acquired" as a result of the "Meeting of Frontiers" partnerships is an open one – at present these digital files live outside of the stream of materials otherwise acquired by the acquisitions units of the Library of Congress – materials that are sent for traditional cataloging or other description and which have some kind of systematic storage to insure the materials' preservation and that there will be access. There is access to the files on the "Meeting of Frontiers" web site but no set plan for how these files become part of LC's "true" collections – an issue that flows from this kind of cooperation that remains to be settled.)

The Library of Congress also has experience working with other institutions – however not as cooperation but rather in the form of a competition – the Ameritech funded competitions allowed the digitization of many collections at U.S. institutions. Again, however, the nature of the competition model has significant difference from a cooperative partnership and allows the Library far more control.

Conclusion

This is a brief overview of some of the more problematic aspects of this cooperative effort. The Library of Congress is excited to have this opportunity work with Russian institutions – we believe this is the only large scale of this type now in operation. Clearly the reason this project has moved forward is because of the flexibility of both LC and the Russian participants – a willingness to adapt as necessary due to circumstances, to compromise as needed, and to avoid getting caught up with overly dogmatic views.

Bibliography

http://frontiers.loc.gov/ - the virtual domain of the "Meeting of Frontiers" web site.

http://frontiers.loc.gov/intldl/mtfhtml/mfprtnrs/mfprtnrs.html – information about the "Meeting of Frontiers" partners.

http://frontiers.loc.gov/intldl/mtfhtml/mfabout/mfdigitz.html – information about certain technical aspects of building the "Meeting of Frontiers" site, note for example the fourth paragraph in which the lack of commonality between the LC records and those supplied by Russian partners is described.

http://memory.loc.gov/ammem/ftpfiles.html – Entry point within the American Memory site to technical information about the building of digital libraries as practiced at the Library of Congress.

http://lcweb2.loc.gov/ammem/award/docs/stepsdig.html – a single page summary, "Steps in the Digitization Process," by Carl Fleischhauer, January 1996.

http://memory.loc.gov/ammem/ndlpedu/index.html - The American Memory "Learning Page" that offers insights to teachers and other educators on how to use American Memory materials in teaching school children.

http://www.loc.gov/today/pr/1999/99-185.html – Original "Meeting of Frontiers" press release.

http://www.loc.gov/today/pr/2001/01-011.html – Press Release about the update to "Meeting of Frontiers" that includes Russian materials.

http://lcweb2.loc.gov:8081/ammem/mtfhtml/archive/granchal.htm – Russian language description of activities of the National Digital Library Program.

http://lcweb2.loc.gov:8081/ammem/mtfhtml/archive/grankach.html – description of quality of files for "Meeting of Frontiers" - in Russian.