

The Use of TEI and OAI in Manuscript, an Informational-Analytical System

Pavel A. Votincev (pvl@udsu.ru)

Udmurtia State University, Izhevsk, Russia

Currently there are at least a few hundred ancient Slavic manuscripts of the 11th to 14th centuries, which, despite their significant scholarly, cultural, and historic value, are not accessible to a wide range of researchers of various specialties and to all who are interested in Russian history and culture. The reason is simple: a large portion of the manuscripts are still unpublished, and if they are published, these editions are rareties, accessible only in major research libraries, which are often located far from scholars' places of work. Even publishing a manuscript often requires constant consultation of the original, especially if studying its paleography, appearance, or other similar questions. Furthermore, existing print editions unfortunately usually lack reference material necessary for getting acquainted with a text and studying it, and the manuscripts themselves are published with poor quality. The solution to the inaccessibility of ancient and medieval manuscripts for research, teaching, and popular use is the creation of electronic publications capable not only of replacing print editions but also of exceeding their capabilities: (a) by providing continual access to digital copies of manuscripts and reference material on the Internet; (b) by providing the user with the capability of retrieving data from a multifunctional interface not only on the level of meta- and analytical data but also on the level of grammatical, semantic, and thematic characteristics of texts; and (c) by allowing collaborative work on a manuscript.

One example of an information system designed to solve some of the above problems is Manuscript, an informational-analytical system.

Manuscript is oriented toward the entry, storage, automatic manipulation, research, and publication of ancient texts having an identical graphico-orthographic appearance and structure to the original. Currently Manuscript contains a few dozen unique ancient and medieval Slavic texts.

Currently Manuscript consists of a few multifunctional modules that interact with a single database containing a hierarchically structured data:

- a specialized editor,

- a query and retrieval module,
- a module for electronic publishing,
- a module of linguistic dictionaries, and
- a module for loading texts and reference material for print publication.

The main task right now of the team behind Manuscript is to present the accumulated information for analysis and study by a wide range of researchers. The main path: to create specialized modules intended for multi-function, user-friendly, and straightforward manipulation of the data. The second task is to create the means to exchange meta-, analytical, and textual data between teams of researchers creating analogous collections of material, with the goal of presenting users with the ability to use manuscript- and text-specific tools for working with ancient documents, created by various teams.

It's essential to use standard formats and protocols to maximize access to the various tools for aggregating information in Manuscript.

The Open Archive Initiative (OAI) and Text Encoding Initiative (TEI) were chosen as the basis for solving this task:

- OAI – for indexing the existing texts and their fragments;
- TEI – for presenting meta- and analytical information, and likewise the texts themselves.

Manuscript is currently working to become an OAI data provider. A new module is being developed to present the metadata in Dublin Core format. This module will allow the gathering of metadata not only from the texts themselves but also from their fragments (various elements of the hierarchy).

Users can retrieve metadata or full text from Manuscript in TEI XML format using a unique identifier.

The TEI Guidelines are well-established around the world, so it was entirely logical for the developers of Manuscript to use it for the metadata and full text of the manuscripts.

Manuscript uses a network model of data for describing the objects (entities) of manuscripts and the relations between them, so the main problem with using TEI was XML's requirement of a strict hierarchy. One consequence of this is that it's impossible to receive all information about a text in Manuscript in one document.

Two modules are being developed to integrate the technology of Manuscript with the TEI.

The module for storage and manipulation of meta- and analytical data is intended for working with archeographical, textological, and other findings from the manuscripts, texts, and fragments themselves, and also for formulating queries and retrieving

information using these findings. Searching in this module is possible only using meta- and analytical data.

A special module processes user queries for retrieving certain elements of text in TEI format from Manuscript. The main capabilities of the module are the following:

- retrieving of any element of text, stored in Manuscript as an independent unit, in TEI format;
- setting priorities in the creation of a TEI document: This option is necessary in order to get around the single-hierarchy limitation of XML. The user can rank the priority of various structures within a document;
- receiving meta- and analytical information with or without full text;
- The queried document is created at the moment of query, so any change will be available immediately.

Manuscript and its tools are being continually developed. During the presentation, Manuscript, including the above modules and technologies, will be demonstrated.

Translated from the Russian by Kevin S. Hawkins.