Editor of Ancient Texts as Part of the System «Manuscript»

Victor Arkadievich Baranov, Andrey Anatolievich Votintsev, Roman Michailovich Gnutikov

Laboratory of Computer-Aided Philological Research, Udmurtia State University
Universitetskaya Str. 1, 426034 Izhevsk, Russia
e-mail: baranov@uni.udm.ru; romashka@uni.udm.ru

Abstract

The Information Retrieval System "Manuscript" is intended for storing, editing and processing electronic copies of manuscripts. By retaining all the peculiarities of the ancient treasures, the Manuscript system provides a thorough input of texts/manuscripts under study while preserving the integrity of the original electronic copy of the manuscript, text transcription, and transliteration for the purpose of creating reference materials or electronic and printed publications of the original manuscripts. The Manuscript system provides a correct input of texts/manuscripts under study by retaining all essential peculiarities of the manuscripts. Computer-aided processing of manuscripts is done taking into account formal and implicit properties of texts and their fragments. The system allows creating text transcriptions and transliterations, preparing dictionaries, lists of words and other units of interest (syntagmas, fragments) for digital and printed editions. The manuscript digital copy exists as a typed electronic text. A specialized text editor "OldEd" was created to ensure data input and correction directly in the database. It helps the user to work effectively with visualized parts of manuscripts, relationships, their properties and values. The editor represents the edited text in the form of geometrical, linguistic, functional and others hierarchies. In the mode of operation over the text, a text can be edited and divided into fragments. In the mode of viewing the hierarchies, the editor allows creating new units (including texts), viewing, adding and deleting relationships between text and dictionary units and modifying their properties (see http://io.udsu.ru/pub/rd/ for more detail).

Keywords: ancient; manuscripts; full-text; databases

Introduction

In operation over ancient texts that are unique by their cultural value a comprehensive study of them becomes very important. For example, in linguistic analysis of ancient and medieval manuscripts consideration of the composition and structure of the manuscripts and texts making part of them and the authorship and genre of the latter, the translation and unique character of the parts and many other factors proves very important for getting objective conclusions. In this case textual studies are an indispensable stage in study of hand-written treasures. In practice the use of already available information by every following investigator begins from the division of the manuscript and retrieved data of the analyzed units according to the available conclusions and every new turn of the investigation requires regrouping data depending on the necessity of taking into account or ignoring these or those text parameters. In any case the practical application of the available textological data is a rather labour consuming work. The way out from this situation seems to be in the availability at investigators' disposal of a certain prototype of the manuscript where diverse and versatile structured information is created and stored and can be easily applied to various investigations.

To ensure an integrated and many-sided investigation of unique hand-written treasures and continuity in the work of various specialists, the text (manuscript) should exist in the form maximum close to the original as a set of minimal structural components (symbols) that are grouped into larger units: word forms, fragments, sections etc. (sometimes of a non-linear composition). These components are assigned values proper to them and by analogy with which other units that are absent in the text, but connected with the existing units (for example, protofragments relative to which text fragments can be studied from the point of view of variant readings, normalized grammar or modern text equivalent etc.), can be created. The above can be realized with the help of full-text databases. It is clear that during all process of work the electronic image of the manuscript should have sure access whatever time it is done and wherever text is stored. In addition, the user work with the full-text database can be valuable and effective only if input, editing, fragmentation, assignment of values, retrieval, grouping and ordering of the textual information is done with the help of a friendly interface that provides the user with the usual manuscript image without dependence on the complexity of the text structure and relationships between units.

Ideology and Theoretical Approaches

The thorough study of texts leads to the revelation of dissimilar relationships between units that are a net when considered in the aggregate. The net is a complicated structure that is inconvenient to operate from the standpoint of organization of user interfaces. This is why we divided the net of units into subnets that are hierarchies processed by the editor. By now the editor operates with the following hierarchies:

- *Geometrical* - The manuscript comprises sheets, sheets comprise pages, pages lines and lines symbols;
- *Linguistic* - The text comprises syntagmas, word forms and symbols; syntagmas comprise word forms and symbols; word forms comprise symbols;
- *The hierarchy of transformed word forms* - Each unique word form of the text is connected with the respective transformed word forms: transformed, normalized, modern etc.
- *Functional-structural* - The text comprises some sections that contain smaller subsections and symbols. For example: the canon consists of a title and songs; the song is divided into a title, tropar' (one or many), canon, bogorodichen and irmos.

The editor allows representation of many hierarchies and also the text in the «flat» form that is a transformed geometrical hierarchy. In the «flat» mode the editor allows adding symbols to the text, deleting them, dividing into lines, pages, sheets. In the mode of operation over hierarchies the editor reflects the information on the relationships between the text units, structural and dictionary units; allows establishing or deleting relationships between units; viewing their properties and creating new units.

Technical Implementation

For operation over the data stored in the full-text databases it makes sense to create special editors. The development and creation of such an editor is now being carried out by the group of developers of the Laboratory of Computer-Aided Philological Research of Udmurtia State University, Izhevsk, Russia. The editor is a module of the system «Manuscript». It provides access to the databases. The key requirements for the editor can be the following:

- Input and editing of the text and additional information on it at the direct interaction with the database;
- Display of the manuscript text in the view close to the original;
- Separation and creation of text units and operations over their properties and values;
- Operation over unit relationships (creation, re-subordination, deleting, change of properties and types of relationships, operation over hierarchical structures);

The above requirements are met by the product that is being created. Today the prototype of the editor where the below listed functions and modes are realized is on the stage of trial operation:

- Multi-user operation at the direct interaction with the database and storage of the results of operations over documents in it;
- Creation of new documents (input, editing and other standard operations of text processors) in the form close to the original;
- Separation of text/manuscript units, entering them into the database and operations over their properties and values;
- Creation and deletion of relationships of various types between units;
- Operations over various hierarchical structures (separation of a unit, creation of a relationship with the parent unit etc.) and representation of the structure of unit relationships as a tree;
- Simultaneous operation over many hierarchies;
- UNICODE- Support.

The access to the database is organized with the help of ADO (ActiveX Data Objects). This allows an independence from the source of data sufficient, for example, to get access to various versions of the Oracle DBMS (Oracle 8, Oracle 9), files XML. ADO also supports the work model 'briefcase', which implies operation with the database without any support of the continuous connection.