OLD SLAVIC MANUSCRIPT HERITAGE: ELECTRONIC PUBLICATIONS AND FULL-TEXT DATABASES

Victor BARANOV, Andrey VOTINTSEV, Roman GNUTIKOV, Aleksey MIRONOV, Sergey OSHCHEPKOV, Vitaliy ROMANENKO
Laboratory of Computer-aided Philological Research
Udmurtia State University
1 Universitetskaya Str., 426034 Izhevsk, RUSSIA
Tel.: +7 3412 363239, Fax: +7 3412 754649
Email: baranov@udm.ru

Abstract
This work covers problems of publication of old manuscripts on the internet and principles of creation of full-text databases for a comprehensive investigation. The presented approach is based on adequacy of the electronic copy to the original, fragmentation of texts depending on the tasks of work, establishment and storage of relationships among the objects under study, use of object attributes for retrieval data. The developed technologies provide a multiuser access by means of internet to the database for data input, editing, processing and retrieval, and creation of scientific, reference, and popular electronic and printed editions of unique manuscripts in any language.

INTRODUCTION

Today several hundred rare and unpublished Slavonic manuscripts of the XI-XIV centuries, which are of great scientific, cultural and historical value, are not accessible to wide circles of researchers or those curious about Slavonic history and culture.

The Information Retrieval System “Manuscript” (Manuscript) that is being developed at the Laboratory of Computer-aided Philological Research, Udmurtia State University is intended for storing, editing and processing electronic copies of manuscripts. The goals of the project are publication of manuscripts (either on the internet (see Fig. 1) or as printed editions [1]) and their use for the textual, linguistic, literary, historical, and cultural investigation in those areas of the Humanities where these types of manuscripts and their components are the objects of study [2].

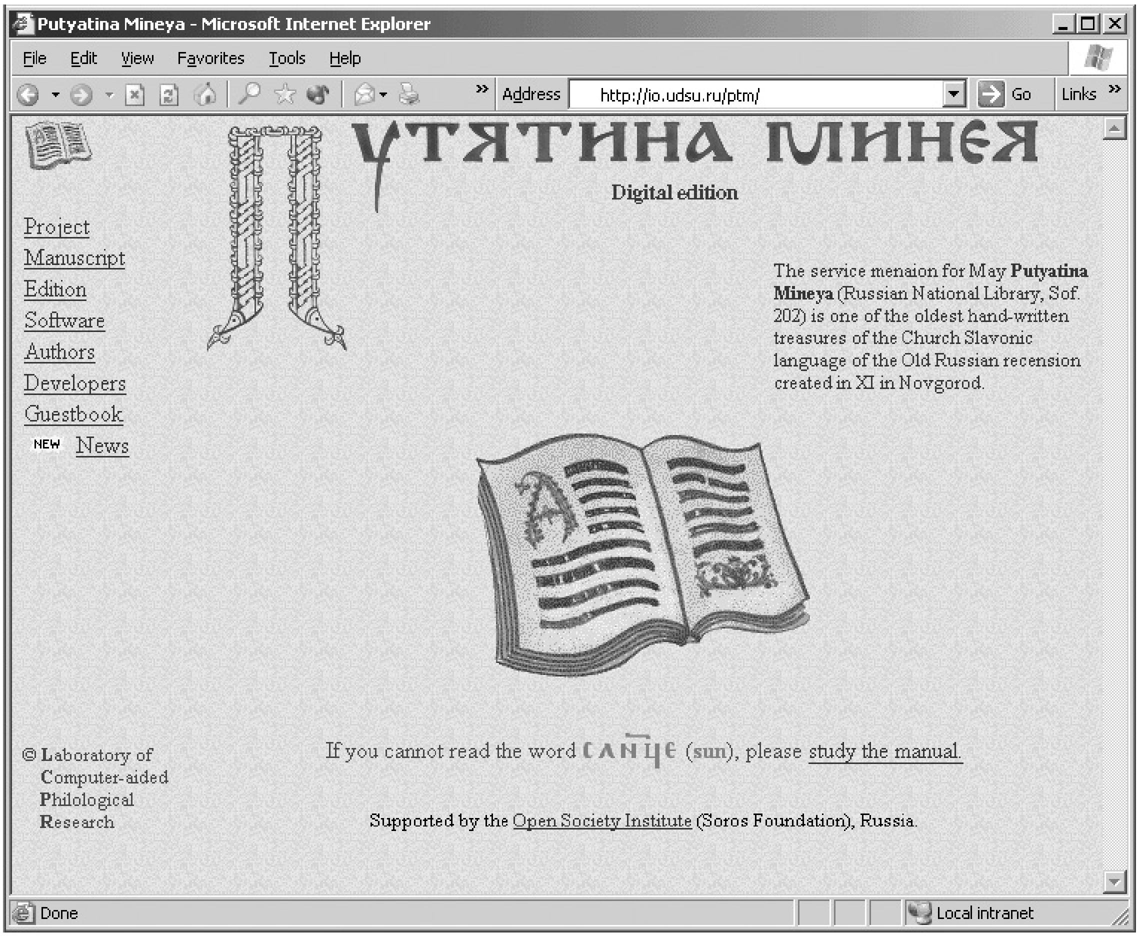
IDEOLOGY AND THEORETICAL APPROACHES

By retaining all the peculiarities of the inscriptions, 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.

The key feature of the Manuscript system is its ability to store and process texts of any graphic–orthographic complexity in any language. Work with manuscripts is based on the following principles:

- Marking out in the text its units of any size: from a fragment to a character;;
- Organization of hierarchies of the above units: character – word-form – syntagma – text, character – line – page – sheet – manuscript, etc.
- Storage of an unlimited list of unit characteristics: values, phonemes, morphemes, words and word-forms, fragments significant from the point of view of textual study, codicology, paleography, archeography, library storage and description.



The service menaion for May **Putyatina Mineya** (Russian National Library, Sof. 202) is one of the oldest hand-written treasures of the Church Slavonic language of the Old Russian recension created in XI in Novgorod.

If you cannot read the word слнце (sun), please study the manual.

© Laboratory of Computer-aided Philological Research

Supported by the Open Society Institute (Soros Foundation), Russia.

Fig. 1. Electronic edition: title page

The main goal of most existing approaches to work with a number of texts is in description of data on text and use of available descriptions for compilation of subject catalogues in library systems. The approach applied in the Manuscript system is in text structuring and use of selected units for creation of structure-subject lists, compilation of alphabetical and quantitative indexes of units. The system offers possibilities for retrieval of the objects of the same type and with the same function (fragments, units) inside texts, which allows obtaining comparative characteristics of not only texts themselves, but also their parts. This helps to reveal stylistic, author’s and other peculiarities of texts and their fragments, for example, for identification of the manuscript authorships.

The practical part of work with text is now carried out on the basis of Old Russian written treasures: service Menaions, service Gospels of the XI-XIV centuries, hagiographies of the XII-XVII centuries and chronicles.

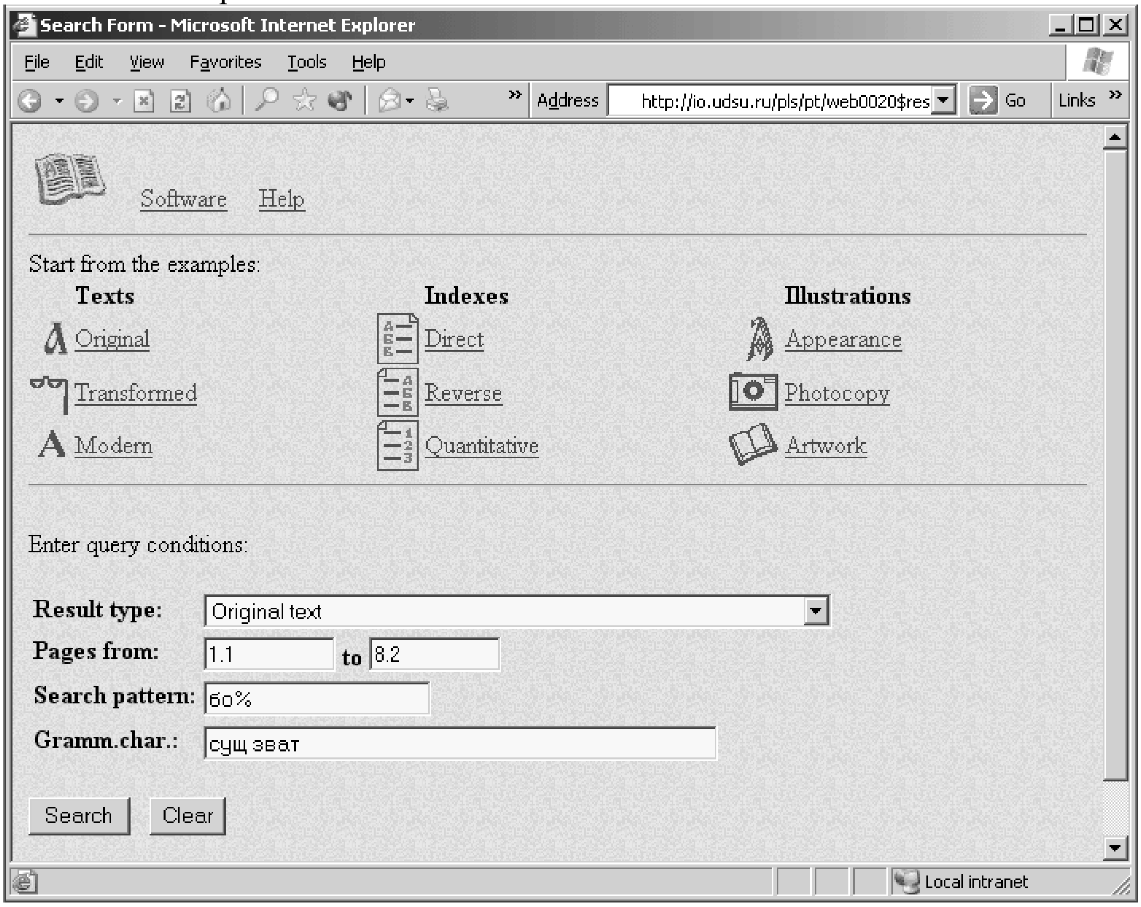
TECHNICAL IMPLEMENTATION.

1. *The Manuscript system.* The information retrieval system "Manuscript" is a complex of programs comprising:

1) database of structurized text information;

2) specialized editor for input and editing of texts stored in the database;

3) search engines presenting texts in various forms and indexes (see Figs 2 and 3), and some other specialized modules.



http://io.udsu.ru/pls/pt/web0020$res

Start from the examples:

Texts	Indexes	Illustrations
Original	Direct	Appearance
Transformed	Reverse	Photocopy
Modern	Quantitative	Artwork

Enter query conditions:

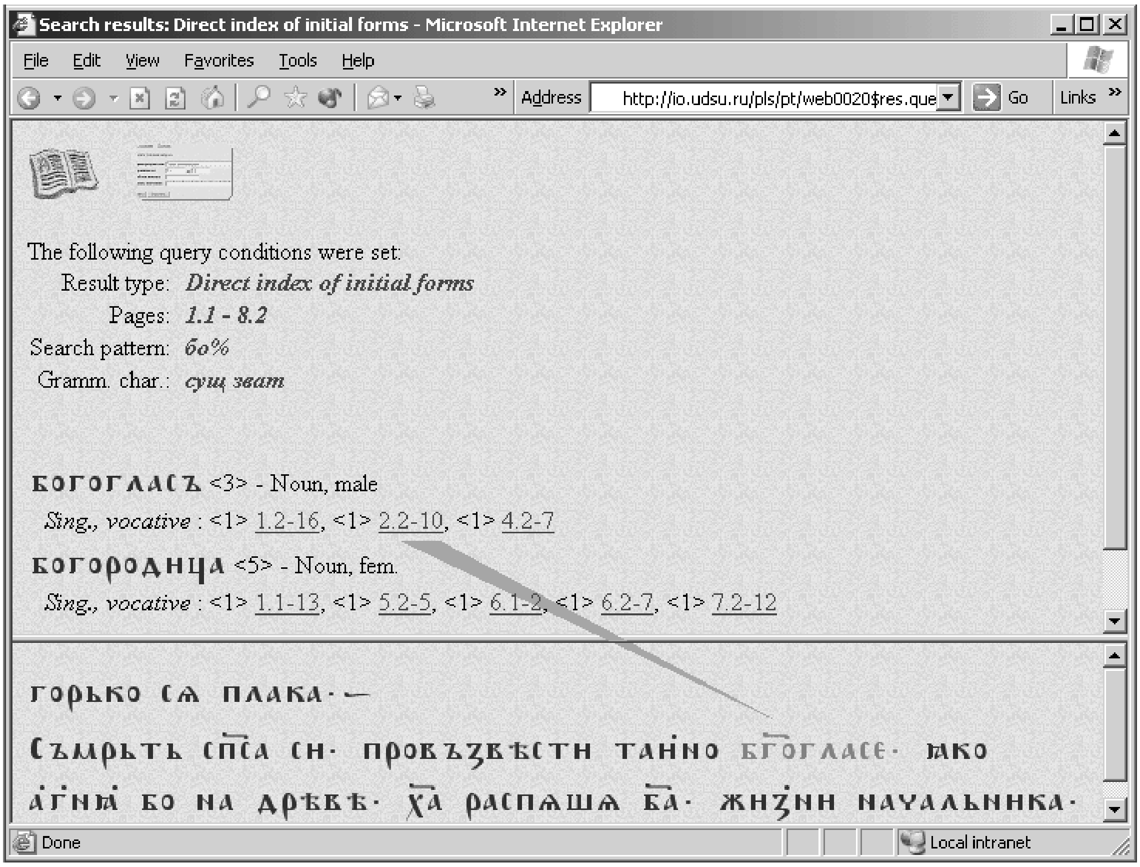
Result type: Original text

Pages from: 1.1 **to** 8.2

Search pattern: бо%

Gramm.char.: сущ зват

Fig. 2. Electronic edition: search form


Search results: Direct index of initial forms - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://io.udsu.ru/pls/pt/web0020$res.que
Go
Links

The following query conditions were set:
Result type: *Direct index of initial forms*
Pages: *1.1 - 8.2*
Search pattern: *бо%*
Gramm. char.: *сущ зват*

богогласъ <3> - Noun, male
Sing., vocative : <1> 1.2-16, <1> 2.2-10, <1> 4.2-7

богородица <5> - Noun, fem.
Sing., vocative : <1> 1.1-13, <1> 5.2-5, <1> 6.1-2, <1> 6.2-7, <1> 7.2-12

горько сѧ плака· —
Съмрьть сп҃са си· провъзвѣсти таино б҃гогласе· ꙗко
аг҃нꙗ бо на дрѣвѣ· х҃а распѧшѧ б҃а· жизни начальника·

Fig. 3. Electronic edition: index of words and word-forms with a concordance fragment.

2. *Units*. The text is divided into units of various types. As known, many old texts do not have any formal division into words and sentences. This is why the Manuscript system helps to find the necessary linguistic units in the text. The user can work not only with linguistic units, but also select fragments written in various time periods, by different authors and with different genre and stylistic characteristics. The list of objects studied in the system can be supplemented with new types of units.

3. *Relationships*. The selected units can be linked by the relationships of various types. The possible types of relationships between units are determined by the user according to the formulated research problems. The relationship between two types of units is determined by its cardinality (one-to-one or one-to-many) and unit types on both ends. The following are the examples of such relationships: inclusion, succession, relationship with dictionary elements.

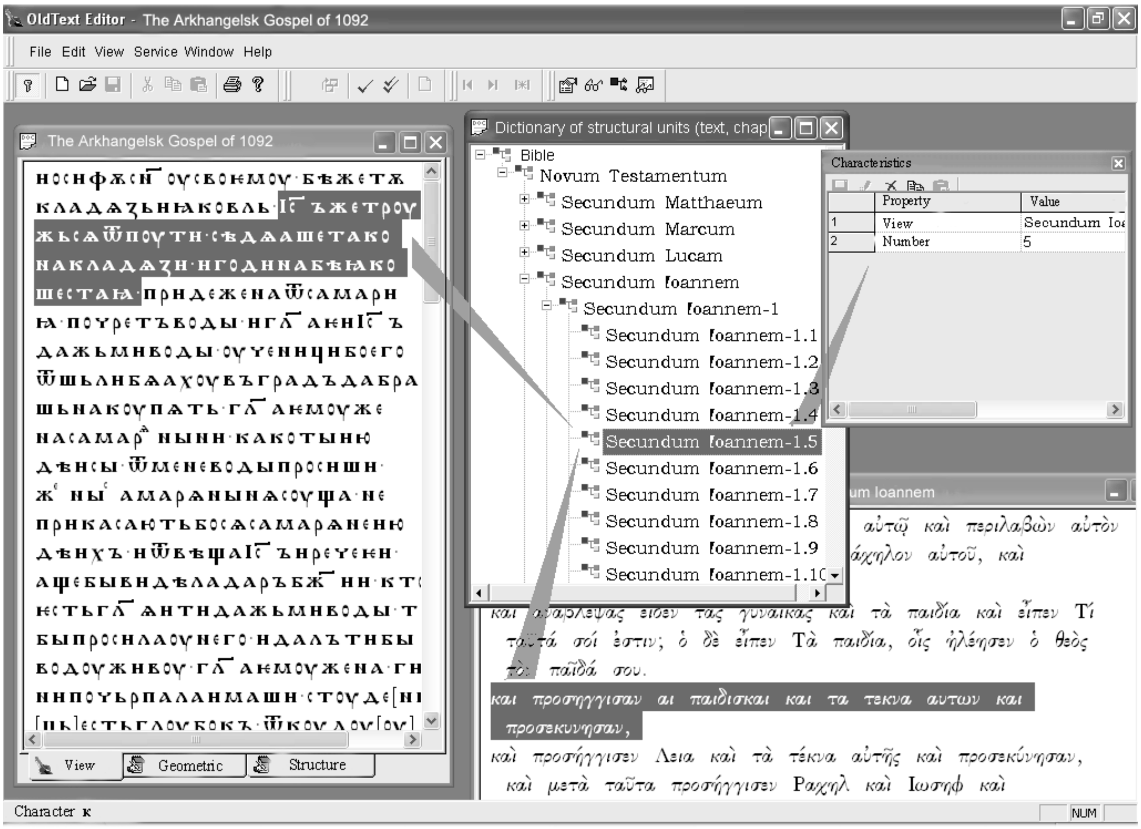
4. *Dictionaries*. The system provides the possibility of maintaining various dictionaries. Any unit can be linked to the dictionary item. The dictionaries considerably decrease the labor expenditures associated with description of units and work with them. For example, the unit “word-form” can inherit some characteristics of the unit “element of the word dictionary”. The dictionaries make possible the examination of texts in term of invariants of linguistic and textual units (annual records of chronics, chapters and verses of Gospel texts etc.). The invariant can exist both in the form of description of identical properties and values of text variants and in the form of reconstructed invariant (i.e. archetype). The latter makes possible revelation of variant readings both within a manuscript, if the repeated units exist, and between the manuscripts containing similar fragments.

5. *Encoding and fonts*. The necessity of storing specific characters of old Slavonic languages and included fragments in other languages in the multi-text database led to the necessity of creation of a specific character set in compliance with the Unicode standard.

The Manuscript encoding system comprises the UTF8LAPREXT1 database character set and the family of Menaion fonts. The encoding system ensures storage, classification and the necessary transformations of all characters used in the multi-textual database. The proposed system of classification of character variants allows relating the character to one of the basic groups of the encoding system, determining its code value and, finally, determining the specific font in which this particular character should be represented.

6. *Specialized editor*. A specialized text editor "Manuscript" 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 [3].

The editor represents the edited text in the form of geometrical, linguistic, functional and others hierarchies. In the mode of operation with text (see Fig. 4), a text can be edited and divided into fragments. In the mode of viewing the hierarchies (see Fig. 5), the editor allows creating new units (including texts), viewing, adding and deleting relationships between textual and dictionary units, and modifying their properties (see http://io.udsu.ru/pub/rd/ for more details).


OldText Editor - The Arkhangelsk Gospel of 1092
File Edit View Service Window Help
The Arkhangelsk Gospel of 1092
View
Geometric
Structure
Dictionary of structural units (text, chap
Characteristics
um Ioannem
Character к
NUM

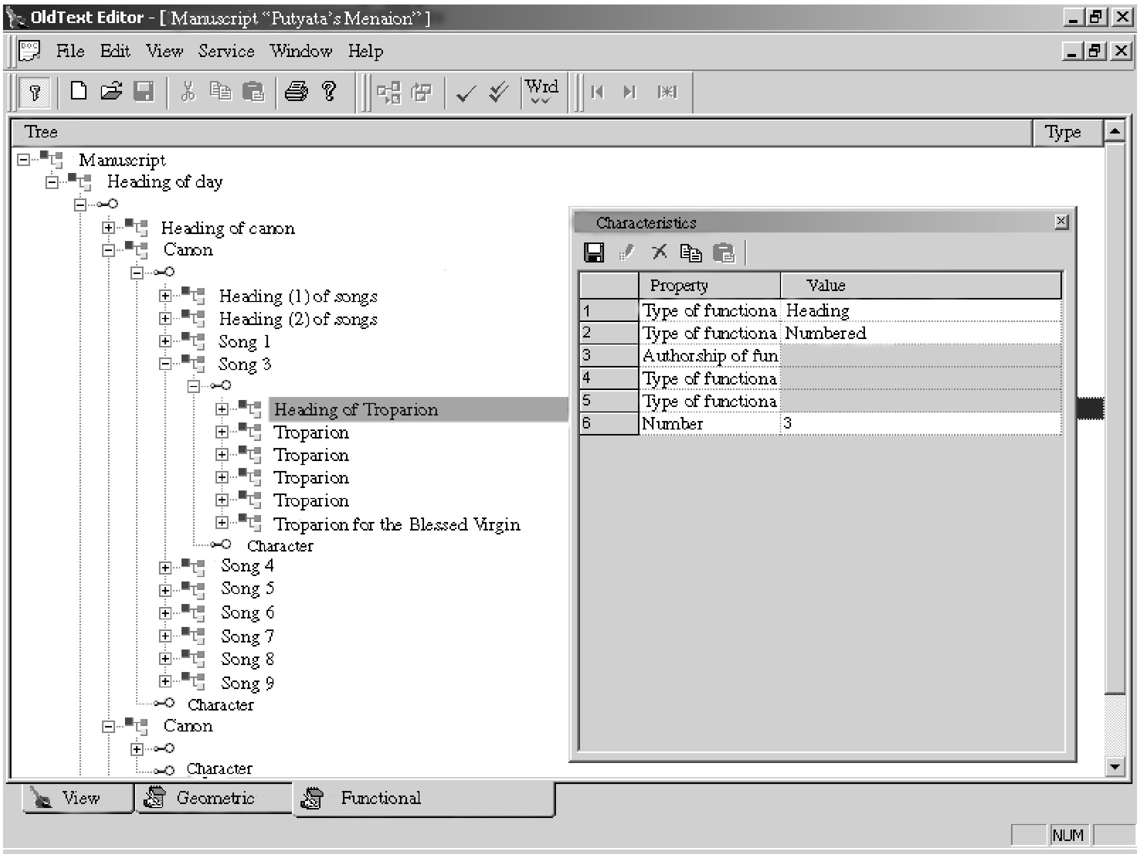
носифѫсн҃ оусвоѥмоу·бѣжетѫ
кладѧзьнꙗковль·Іс҃ ъжетроу
жьсѧѿпоутн·сѣдѧашетако
накладѧзн·нгодннабѣꙗко
шестаꙗ·прндеженаѿсамарн
ꙗ·поѵретъводы·нгл҃ аѥнІс҃ ъ
дажьмнводы·оуѵеннцнбоѥго
ѿшьлнбѣахоувъградъдабра
шьнакоупать·гл҃ аѥмоуже
насамарⷶ нынн·какотынѭ
дѣнсы·ѿменеводыпросншн·
жⷭ ныⷭ амаранынасоуща·не
прнкасаютьбосасамараненю
дѣнхъ·нѿвѣщаІс҃ ънреѵеѥн·
ащебывндѣладаръбж҃ нн·кт
ѥстьгл҃ антндажьмнводы·т
быпроснлаоунего·ндалътнбы
водоужнвоу·гл҃ аѥмоужена·гн
ннпоѵьрпаланмашн·стоудє[нн
[нь]естьглоубокъ·ѿкоудоу[оу]

Bible
Novum Testamentum
Secundum Matthaeum
Secundum Marcum
Secundum Lucam
Secundum Ioannem
Secundum Ioannem-1
Secundum Ioannem-1.1
Secundum Ioannem-1.2
Secundum Ioannem-1.3
Secundum Ioannem-1.4
Secundum Ioannem-1.5
Secundum Ioannem-1.6
Secundum Ioannem-1.7
Secundum Ioannem-1.8
Secundum Ioannem-1.9
Secundum Ioannem-1.10

	Property	Value
1	View	Secundum Ioa
2	Number	5

αὐτῷ καὶ περιλαβὼν αὐτὸν
άχηλον αὐτοῦ, καὶ
καὶ ἀναβλέψας εἶδεν τὰς γυναῖκας καὶ τὰ παιδία καὶ εἶπεν Τί ταῦτά σοί ἐστιν; ὁ δὲ εἶπεν Τὰ παιδία, οἷς ἠλέησεν ὁ θεὸς τὸν παῖδά σου.
και προσηγγισαν αι παιδισκαι και τα τεκνα αυτων και προσεκυνησαν,
καὶ προσήγγισεν Λεια καὶ τὰ τέκνα αὐτῆς καὶ προσεκύνησαν,
καὶ μετὰ ταῦτα προσήγγισεν Ραχηλ καὶ Ιωσηφ καὶ

Fig. 4. Editor: linking of the selected fragment to the item of the dictionary of fragments



	Property	Value
1	Type of functiona	Heading
2	Type of functiona	Numbered
3	Authorship of fun	
4	Type of functiona	
5	Type of functiona	
6	Number	3

Fig. 5. Editor: text structure and fragment properties

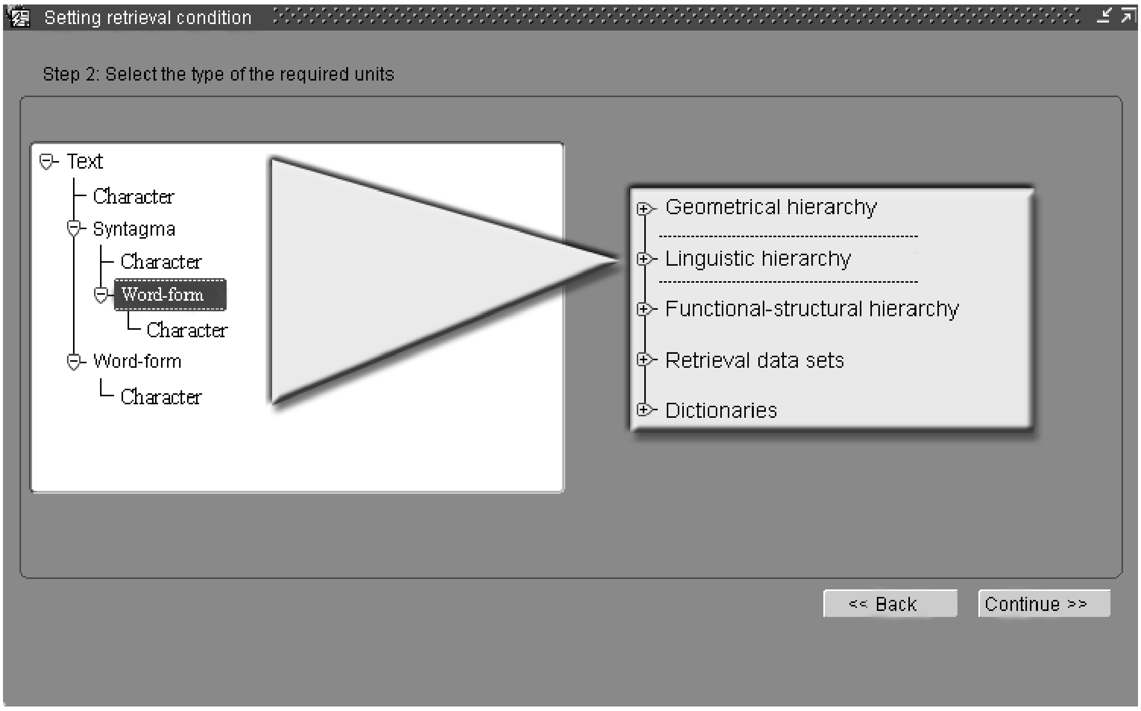
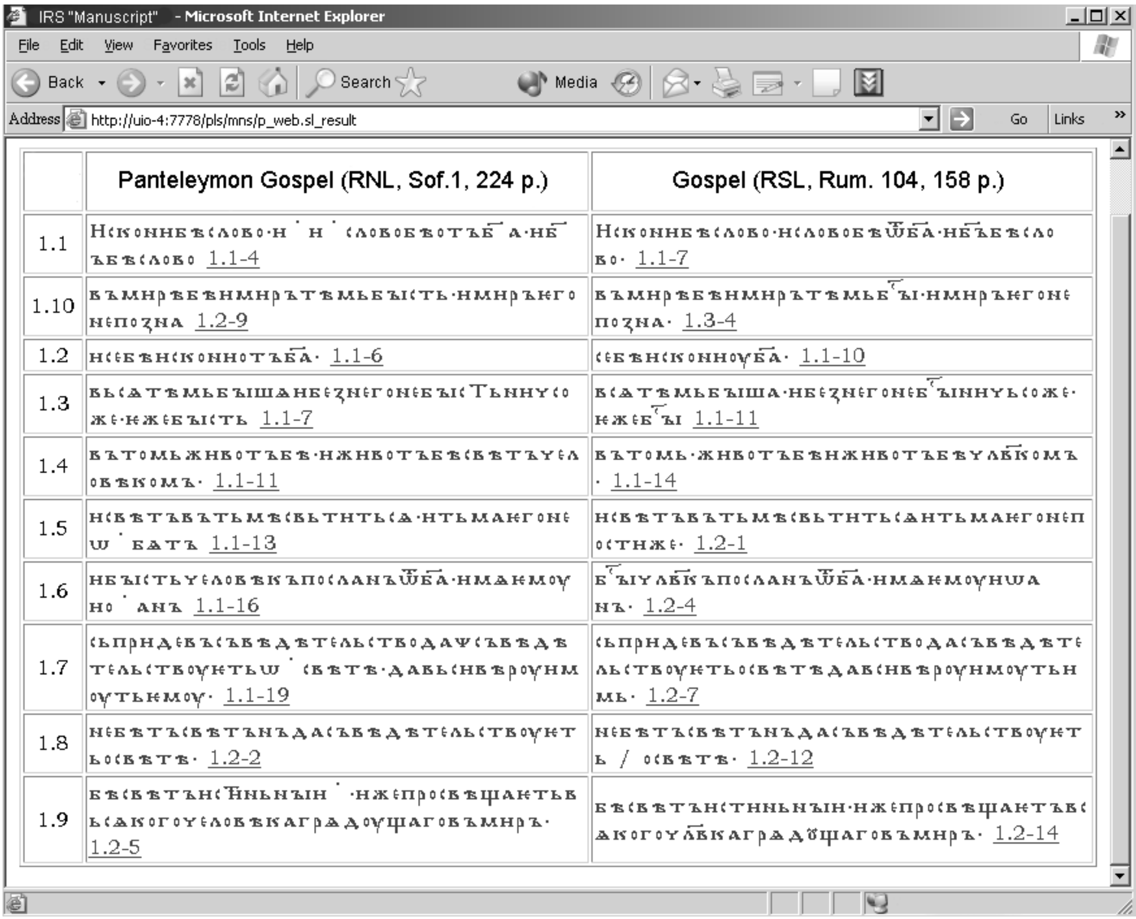


Fig. 6. Retrieval module: selection of a hierarchy and its objects


IRS "Manuscript" - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Media
Address
http://uio-4:7778/pls/mns/p_web.sl_result
Go
Links

	Panteleymon Gospel (RNL, Sof.1, 224 p.)	Gospel (RSL, Rum. 104, 158 p.)
1.1	Исконибѣслово·и ̇ и ̇ словобѣотъб҃а·иб҃ъбѣслово 1.1-4	Исконибѣслово·исловобѣѿб҃а·иб҃ъбѣслово· 1.1-7
1.10	въмирѣбѣимирътѣмьбысть·имиръєгонєпозна 1.2-9	въмирѣбѣимирътѣмьб҃ы·имиръєгонєпозна· 1.3-4
1.2	исєбѣискониотъб҃а· 1.1-6	сєбѣискониоуб҃а· 1.1-10
1.3	вьсатѣмьбышаибєзнєгонєбыстьничсожє·ижєбысть 1.1-7	всатѣмьбыша·ибєзнєгонєб҃ыничьсожє·ижєб҃ы 1.1-11
1.4	вътомьживотъбѣ·иживотъбѣсвѣтъчєловѣкомъ· 1.1-11	вътомь·животъбѣиживотъбѣчл҃вкомъ· 1.1-14
1.5	исвѣтъвътьмѣсвьтитьса·итьмаєгонєѡ ̇ батъ 1.1-13	исвѣтъвътьмѣсвьтитьсаитьмаєгонєпостижє· 1.2-1
1.6	ибыстьчєловѣкъпосланъѿб҃а·имаємоуио ̇ анъ 1.1-16	б҃ычл҃вкъпосланъѿб҃а·имаємоуиѡанъ· 1.2-4
1.7	сьпридєвъсъвѣдѣтєльстводаѱсъвѣдѣтєльствоуєтьѡ ̇ свѣтѣ·давьсивѣроуимоутьимоу· 1.1-19	сьпридєвъсъвѣдѣтєльстводасъвѣдѣтєльствоуєтьосвѣтѣдавьсивѣроуимоутьимь· 1.2-7
1.8	нєбѣтъсвѣтъиъдасъвѣдѣтєльствоуєтьосвѣтѣ· 1.2-2	нєбѣтъсвѣтъиъдасъвѣдѣтєльствоуєть / освѣтѣ· 1.2-12
1.9	бѣсвѣтъист҃иньныи ̇ ·ижєпросвѣщаєтьвьсакогочєловѣкаградоущаговъмиръ· 1.2-5	бѣсвѣтъистиньныи·ижєпросвѣщаєтъвсакогочл҃вкаградꙋщаговъмиръ· 1.2-14

Fig. 7. Retrieval module: comparative list of fragments

7. *Queries and retrieval results.* A specialized retrieval module of the Manuscript system was designed and created as a web application (see Fig. 7). It provides for selection of manuscripts and texts, types of units, setting selection criteria, selection of the presentation type and sorting order of retrieval results as needed (see Fig. 6). The retrieval module allows different operations with the prepared retrieval results (union, intersection and difference).

8. *Multi-user operation*. The Manuscript system enables multi-user access to texts both for data input into the database by means of the Manuscript editor and for querying using single-text search engines.

9. *Electronic publications* of two Old Russian manuscripts can be found on http://io.udsu.ru/ptm_en/ (Putyatina Mineya, XI cent.) and http://io.udsu.ru/pev/ (Panteleymon Gospel, XII-XIII cent.).

10. *Perspectives.* Today creation and further development of means of computer-aided morphological, syntactical and semantic analysis of ancient and medieval Slavonic texts are promising and awaited by specialists involved in text processing and analysis. The development of algorithms and technologies of comparative analysis of texts and their fragments in any languages is also of interests.

The technologies presented in this paper allow storing and processing texts and preparing printed and electronic publications of manuscripts characterized by complex compositions and structures. What is important and should be stressed is that the Manuscript system provides wide possibilities for carrying out complex works necessary both for thorough and deep research and popularization of knowledge on hand-written treasures belonging to the cultures of various nations. Analysis of a number of texts would allow clearing up the questions of the history of texts, their origin and authorship. It would also help to reveal and stress the historical and cultural value of every individual manuscript coming to us through centuries.

ACKNOWLEDGMENT

The work was made possible thanks to the financial support of the Russian Foundation of Basic Research (Grant 02-07-90424).

References

[1] Novgorodskaja sluzhebnaja mineja na maj (Putjatina mineja). XI vek: Tekst, issledovanija, ukazateli (Novgorod May Service Mineya (Putyatin mineya). XI century: Text, investigation, indexes). (2003) Edited by V.A. Baranov, V.M. Markov. Izhevsk, 788 p.

[2] BARANOV, V.A., VOTINTSEV, A.A., GNUTIKOV, R.M., ZUGA, O.V., MIRONOV, A.N., NIKIFOROVA, S.A., OSHCHEPKOV, S.V., ROMANENKO, V.A. and RYABOVA, E.V. (2003) Electronnyje izdanija drevnikh pis’mennykh pamjatnikov i tekhnologija sozdanija polnotekstovykh baz dannykh (Electronic Editions of Old Written Monuments and Technology of Creation of Full-text Databases). Krug idej: electronnye resursy istoricheskoj informatiki, Moscow, pp. 234–260.

[3] BARANOV, V.A., VOTINTSEV, A.A., GNUTIKOV, R.M., MIRONOV, A.N. and ROMANENKO, V.A. (2003) Spetsializirivannyj tekstovyj redactor “Manuscript” Sistemy obrabotki drevnikh rukopisej (Specialized Text Editor Manuscript of the System for Processing Old Manuscripts). Informatsionnyj bjulleten’ assotsiatsii “Istorija i komp’juter 31: 159-165.