Acta et Commentationes, Exact and Natural Sciences
Volume 2(10), 2020, Pages 97–107
ISSN: 2537-6284
E-ISSN: 2587-3644
DOI: 10.36120/2587-3644.v10i2.97-107

# Perspective research at the Vladimir Andrunachievici Institute of Mathematics and Computer Science

Inga Ţiţchiev and Constantin Gaindric

**Abstract.** The article shows the relevance and benefits of research conducted in the field of computer science at the Vladimir Andrunachievici Institute of Mathematics and Computer Science.
**Keywords:** research, ICT, unstructured heterogeneous content, generation of digital content, poorly structured issues.

# Cercetări de perspectivă în cadrul Institutului de Matematică și Informatică ”Vladimir Andrunachievici”

**Rezumat.** În articol se arată relevanța și beneficiile cercetărilor efectuate în domeniul informatic în cadrul Institutului de Matematică și informatică ”Vladimir Andrunachievici”
**Cuvinte cheie:** cercetare, TIC, conținut eterogen nestructurat, generarea conținutului digital, probleme slab structurate.

## 1. Introduction

The massive use of information and communication technologies revolutionizes the development of modern society, consistently contributes to the implementation of the digital society concept. European Digital Agenda proposes the dynamization and optimization of the benefits of information technologies for economic growth, new jobs creation, the improvement of the citizens' quality of life as a part of the Europe 2020 Strategy. The initiative 'A Stronger Digital Europe towards 2025', which brings together about 70 of the strongest European and transnational companies, states in its manifest that digital technologies, innovations and artificial intelligence will create a strong digital Europe, unfragmented, which will lead to the digital inclusion development, increasing the use of 'green' resources, innovations, the development of agile policies, which will ensure prosperity of the European society and will place Europe as a leader in the global economy. The basic research at the institute is carried out in the field of intelligent information

systems with applications in three social domains: medicine, education and culture. In relation to these domains, the conducting research and development of information systems aimed at solving scientific and social problems that are, as a rule, ill-structured, operate with large volume of data, depend greatly on the decision maker's vision and need a personalized approach. This approach involves completion of certain stages of knowledge processing: examination, experiment, conceptualization and analysis, that will serve as a basis for computer science applications in the domains of preservation of cultural heritage, support in medical diagnostics, mitigation management of disasters with multiple victims, automation of the process of design and generation of digital content for computer-assisted learning (e-learning). The proposed solutions will take into account the fragmented and heterogeneous nature of information, data and knowledge in order to define some standardized structures, which will facilitate interoperability and efficient incorporation into the information systems. Regarding the digitization of cultural heritage, the emphasis will be placed on Romanian works printed in the Cyrillic script, covering the period of 17-20 centuries, and having as a result both the printed format with original characters and the transliterated one in the Latin script, adapted to the modern language. For the first time there is addressed the problem of integrated processing of different types of content (text, graphics, formulas, musical notes, etc.). For processing large volume of data, the methods based on formal computational models will be used to ensure parallel processing (including membrane computing and P systems, Petri nets, etc.). Following the purpose of making information processing more efficient, the problems of design and implementation of distributed computing systems will be solved as well, the solutions aimed to improve the functioning of the systems from the point of view of adapting and adjusting the execution environments, taking into account the specific requirements of the various application classes, will be proposed and analysed. Research in order to automate the process of digitization and transliteration of Romanian texts printed in Cyrillic characters began in the Republic of Moldova in 2016 [1] and are today the most advanced in this field. The resulting accuracy is 95The novelty and topicality of the research carried out consists in: a) the new technology developed for the automation of the digitization processes of the poorly structured heterogeneous Romanian texts printed in Cyrillic characters from the 17th-20th centuries and of high volume; b) methods of preprocessing of scanned texts (images) based on formal calculation models, which ensure parallel processing (including membrane calculation) [3]; c) methods of automatic alignment of old texts to the contemporary language; d) creation of a smart Web platform that integrates existing tools and those developed within the project. Computer

applications for the educational process involve the development of specific software tools [13, 14] dedicated to:

- management of the educational process,
- the teaching process,
- construction of educational materials.

The tendencies to ensure a degree of intelligence of information systems are among the most promising for the propulsion of scientific fields and applications of computer science. This will expand the range of problems that can be solved using information systems [4], especially in poorly structured areas, and increase the level of information assistance [5] of a modern specialist in various fields of activity. The relationship between the decision maker's ability to cope with the situation and the degree of difficulty and uncertainty of the problem to be solved is a key moment in the development of an information system. The ideas of creating computer systems that would propose solutions to real problems, taking into account intuition, the vision of the decision maker who requires a personalized approach [6] is observed in the literature, including medicine, since the late twentieth century [7- 11]. Research on decision support systems with behavioral elements began in the Republic of Moldova in 2015 [12]. Thus, the solutions proposed in the project have a perspective in terms of contemporary trends, the needs of society [7] and the possibilities of the project research team. The design and development of information systems taking into account the uncertainty of management objectives, the influence of intuitive factors in decision making, and cause-effect relationships is the innovative approach of research. The efficiency of the solutions will be supported by new and original methods of using high-performance computing technologies and modern infrastructures (GRID, HPC clusters, cloud).

## 2. Objectives pursued

In order to carry out the research, the following objectives were formulated:

- Development of standardized data and knowledge structures from various sources and fields studying the taxonomies / ontologies associated with them and taking into account the fragmentary and heterogeneous structure of information. Ensuring the interoperability and coherence of these structures in order to incorporate them into information systems, including through personalized approaches.
- Development of intelligent IT tools to assist decision makers in solving poorly structured problems, taking into account the fragmentary and heterogeneous structure of data and knowledge.

- Substantiation of the approach in medical diagnosis based on both quantitative characteristics and accumulated knowledge, intuition, reasoning difficult to formalize.
- Mathematical modeling of natural processes, which requires the acquisition, storage and processing of large volumes of data in order to monitor and forecast their evolution.
- Development of the model of a universal platform for processing large volumes of texts with different content, which will contribute to automation of digitization and transliteration of old Romanian texts printed in Cyrillic characters, pre-processing and post-processing of heterogeneous texts, aligning old texts with contemporary ones.
- Development of information systems to automate the process of designing and generating digital content for computer-assisted learning (e-learning) with the involvement of knowledge bases, reusable language resources, modern technologies for programming, processing and visualization of images and large volumes of data.
- Research and development of formal calculation models based on Minsky machines, Petri nets, Chomsky grammars and variants of P systems: transitional, catalytic, antiport, with active membranes, with control, focusing on computing power, efficiency, descriptive complexity.

## 3. The relevance and benefit of prospective research

### 3.1. Research and elaboration of the structure of the software tool for preprocessing the unstructured heterogeneous content

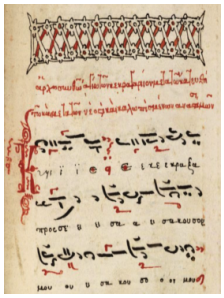
Most documents, in addition to text, also contain other elements: mathematical and chemical formulas, musical notes, diagrams, diagrams, images, etc. The main features of a heterogeneous document are the following:

- the document is not exclusively in natural language;
- there is one or more scripting languages for presenting its components;
- the graphic representation can be rendered through scripting language.

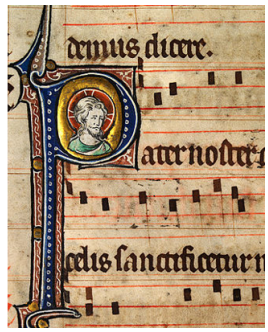
Figure 1.

In Figure 1. the double line separates text of the Hurian Hymn in the upper part and musical scores in the lower part. Additionally, the tablet contains the title of the song and the name of the scribe. Now we call them metadata.

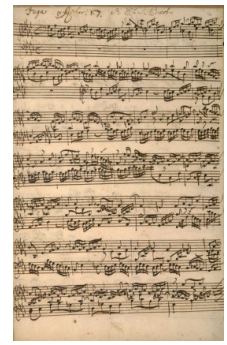
Notation of music in Europe in different historical periods



12th century



14th century



18th century

Figure 2.

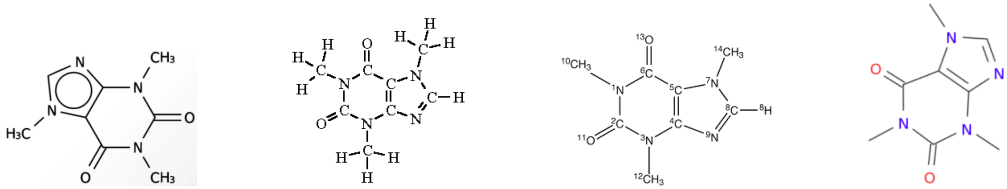


Figure 3.

In Figure 3 the diversity of structure diagrams for the same caffeine is done. Processing issues:

- The problem of processing heterogeneous documents is not completely solved. There are partial solutions for certain types of elements,
- It is difficult, and sometimes even impossible, to automatically recognize certain types of heterogeneous content,
- Page structure analysis is a complex issue,
- A platform is needed to integrate different types of scripts into a unified form of document presentation and processing.

The structure of the semi-automatic workflow for the recognition of heterogeneous documents was proposed in Figure 4.

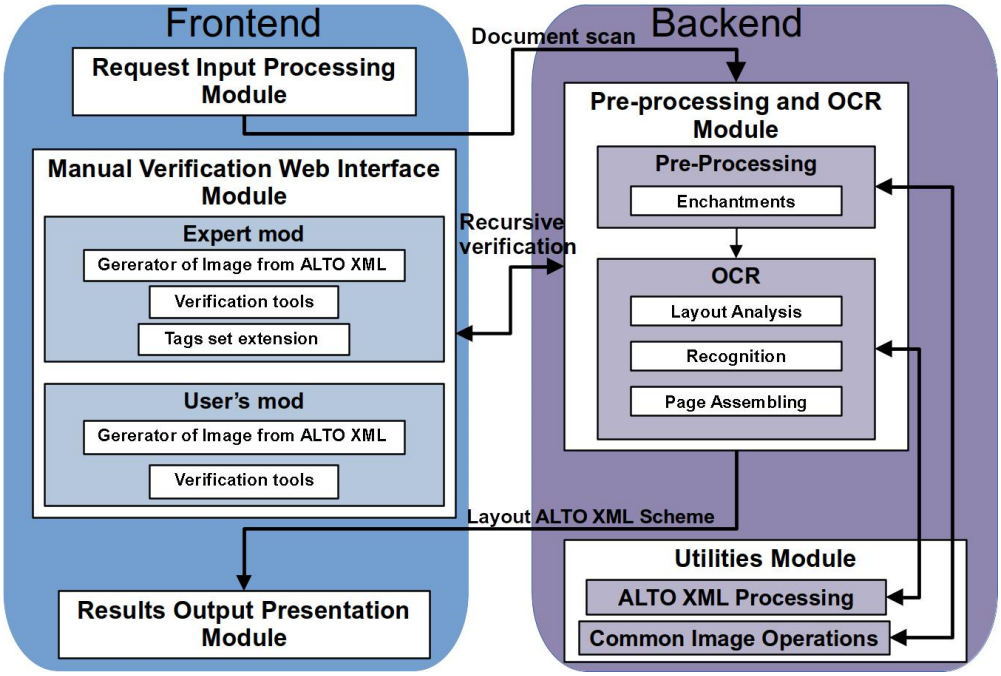



Figure 4.

Automated, semi-automated and manually performed functions in the proposed structure are:

(1) automatize
   (a) scanning;
   (b) recognition of segments according to their type;
   (c) assembling presentations in scripting language with metadata integration;
   (d) rebuilding the page image based on the script;
   (e) automated verification.

(2) semi-automatic
   (a) improving image quality;
   (b) page layout analysis;
   (c) distribution of tasks for manual verification.

(3) manual
   (a) manual verification and correction of the document by experts (for cases where automated or semi-automated procedures could not be applied).

*Partitioning and mapping of heterogeneous documents*

ABBYY FineReader Engine was used, which processes one page of document and returns the result in XML format indicating the coordinates of the page segments and the type of segment (text, image, table, separator, etc.), as well as the recognized text for the

segments that presents fragments of text. A module has been developed in the Phyton language, which:

- Uses Docker container technology as a basis for optimizing platform structure and data flows.
- Extract the metadata from the resulting XML and call a module (Image Magick) for cropping images.
- Organizes the structured storage of results.

The process diagram involves completing the following steps:

(1) Obtaining / retrieving scanned objects in jpeg / tiff format.
(2) Preparing images for OCR (Scan Tailor): cleaning images.
(3) Optical Character Recognition (OCR) with ABBYY Finereader (AFR).
(4) Save the recognized text as a Microsoft Word document (default).
(5) Transliteration of the text obtained in modern Latin script.
(6) Manual / semi-automated processing of the obtained result and placement in the WEB.
(7) Fill in the word lists for AFR (Notepad ++).
(8) Configuring AFR depending on the era, locality and typography.
(9) Configure the virtual keyboard with letters specific to the old alphabets.

Some of the intermediate results are presented in Figure 5.

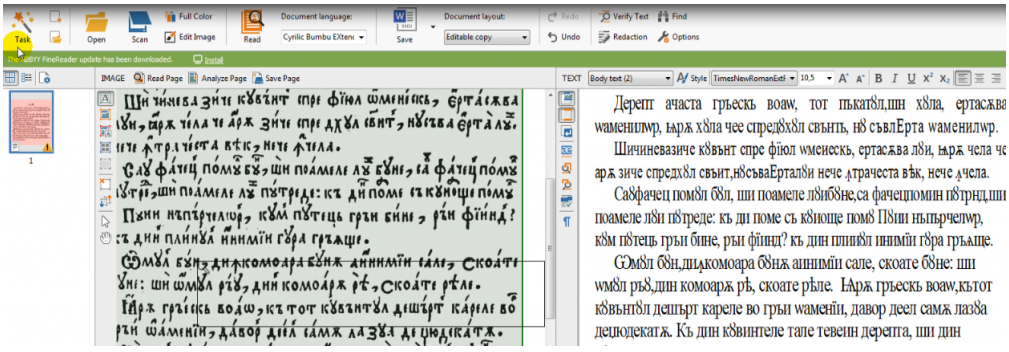


Figure 5.

## 3.2. Research and testing of the possibilities of applying modern programming technologies (object oriented, Web, functional, logical) and information resources to the generation of digital content of computer-assisted training courses.

An instrument for the automatic generation of digital content of computer-assisted training courses on "Finite Automata", using the graphical representation of automata, made with the application of modern functional programming technology and the graphic

editor LaTeX + TikZ was developed. The generated graphic structures are optimized using the Sugiyama framework scheme, in which:

- changes were proposed and made to the “greedy” algorithm for suppressing cycles, taking into account the specifics of the graphical representation of finite automata;
- the stratified algorithm of the acyclic oriented graph by depth traversal was performed;
- effective heuristic methods have been proposed and implemented to reduce the number of arc intersections in a connected and acyclic oriented graph.

### 3.3. Research and development of knowledge and data processing methods for poorly structured issues.

A case study was performed for medical diagnosis as a field with poorly structured and heterogeneous data and knowledge. In particular, the specifics of the field of "multi-victim disasters" were examined.

The information on the accumulation and location of the free liquid was acquired based on the sonographic characteristics and the volume of air accumulated and structured in the form of facts - preparatory action for the formalization of the field and the subsequent creation of the decisional rules.

The information about the victim’s condition was acquired based on vital signs.

The prototype of the formalized domain was created. Its (iterative) validation by expert doctors has begun.

The specific methods of artificial intelligence in the representation of professional knowledge were analyzed in order to identify the most appropriate methods for the field of the research project. In the field of medical examination, the representation of knowledge in the form of taxonomy / tree corresponds to the reasoning of the experts.

The DICOM Network medical image storage and processing system was developed.

An advanced algorithm for archiving medical images using a multi-level data storage system has been proposed. The algorithm takes into account the special architectures of the storage system - one of the main priorities of the algorithm is adaptability to increase system performance, which should provide high-speed access to a huge amount of archived data.

Methods for organizing cloud storage resources for implementing multi-level structures for data archiving are being researched. Methods of operation of different applications with specific parameters and requirements for the structure of data storage organization are analyzed. A method for optimizing the transmission and processing of medical image

data sets has been developed. The main idea of the method is to use virtual machines distributed directly on Cloud Storage to process data.

The proposed solution for the realization of the distributed system of collection, storage, archiving and processing of medical imaging data is presented in Figure 6:

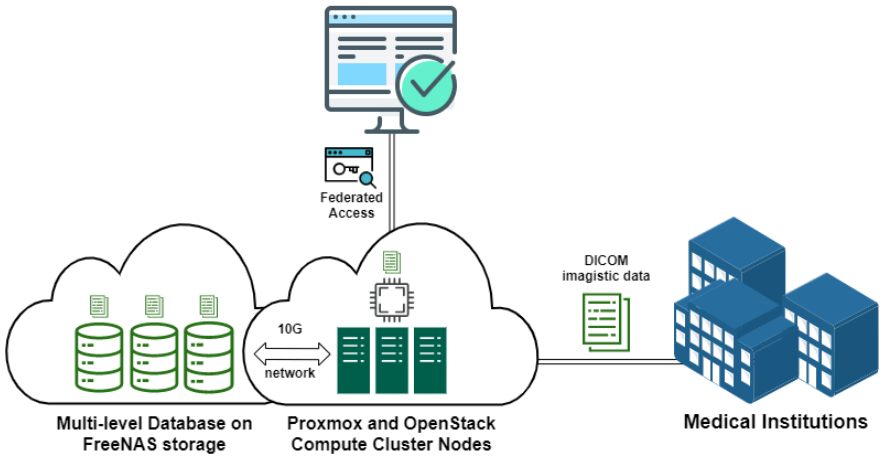



Figure 6.

## 4. Conclusions

The functioning of information systems, as a rule, is fed by the data with which the databases are populated, this popular one requiring overcoming the predominantly unstructured character of the data. Thus, the processing of poorly structured data and knowledge continues to be an important topic. Processing methods currently depend to a large extent on the area examined and an extension is unlikely until new ideas emerge and new progress is made. In order to obtain useful solutions expected from the application of information systems, specific tools are needed to improve the quality of data, information and knowledge.

The research comes with a national contribution to a current challenge of enabling online access to European heritage digital resources. The digital resources of European heritage are potentially significant for the cultural and creative economy sectors, which today are key directions for achieving economic benefits. The results of the research will bring their contribution in facilitating the reprinting of books and other old prints, they are already requested by libraries, archives, researchers in various fields. They will substantially expand the circle of people, who will be able to freely use this component of the national heritage. Carrying out these works would allow the unification, homogenization

and integration of the national-cultural environment in the international information society, would confirm the status of the Romanian language as a language of communication on the European continent.

The knowledge-based society and economy involve the intensive use of information and communication technologies in all spheres of human activity, including in the educational process. Current educational practice involves the development and use of intelligent information systems in order to increase the diversity and quality of computer-assisted training (e-learning). The results will facilitate the generation of digital content of computer-assisted training courses with the application of knowledge bases, reusable language resources, modern technologies for programming, image processing and visualization, large volumes of data.

The systemic treatment of the PSS solution, taking into account the decision-maker's vision and selecting the solution based on a personalized end-user approach is a new way to solve the PSS, the importance and necessity of which is obvious.

Formal computational models will contribute to the development of parallel algorithms for solving a series of difficult problems in linguistics, biology, computational algebra, high performance computing, etc.

In the development of intelligent information systems for various fields, the engineering approach is usually applied, and the modeling of human intelligence is a fundamental concern that influences practical applications.

## References

[1] S.COJOCARU, A.COLESNICOV, L.MALAHOV. Digitization of Old Romanian Texts Printed in the Cyrillic Script. Proceedings of Second International Conference on Digital Access to Textual Cultural Heritage, DATeCH2017, Göttingen, Germany, June 1-2, 2017, ACM, 2017, pp.143-148, Editors: Apostolos Antonacopoulos, Marco Büchler. http://dl.acm.org/citation.cfm?id=3078093

[2] MARIA MORUZ, ADRIAN IFTENE, MIHAI ALEX MORUZ, DAN CRISTEA. Automatic alignment of old Romanian words using lexicons. In A. Moruz, et al. (eds.). Proceedings of the 8th International Conference "Linguistic Resources And Tools For Processing Of The Romanian Language", 8-9 December 2011, 26-27 April 2012,Bucharest, Romania, „Alexandru Ioan Cuza" University Publishing House, Iași, pages 119-126.

[3] GH. PĂUN. Membrane Computing- An Introduction, Springer, 2002, 419 p.

[4] ISABELLE BICHINDARITZ, SACHINVAIDYA,ASHLESHA JAIN, AND LAKHMI C. JAIN (Eds.) Computational Intelligence in Healthcare, Springer, 2010, 470 p.

[5] MORRIS F. COLLEN, KATHRYN J. HANNAH, MARION J. BALL (Series Editors), Computer Medical Databases, ISBN 978-0-85729-961-1, Springer, 2012, 270 p.

[6] TVERSKY, AMOS, and DANIEL KAHNEMAN. Advances in Prospect Theory: Cumulative Representation of Uncertainty, Journal of Risk and Uncertainty, 1992, 5, pp. 297—323.

[7] BARBERIS, N. C. Thirty years of prospect theory in economics: A review and assessment. Journal of Economic Perspectives, 2013, 27(1), pp.173-196.

[8] RICCARDO BELLAZZI, FULVIA FERRAZZI and LUCIA SACCHI. Predictive data mining in clinical medicine: a focus on selected methods and applications, Advanced Review, January / February, 2011.

[9] SUNG HO HA and SEONG HYEON JOO. A Hybrid Data Mining Method for the Medical Classification of Chest PainInternational Journal of Computer and Information Engineering 4:1 2010.

[10] LORIS NANNI, SHERYL BRAHNAM, ALESSANDRA LUMINI, AND TONYA BARRIER, Data Mining Based on Intelligent Systems for Decision Support Systems in Healthcare. In: Advanced Computational Intelligence Paradigms in Healthcare 5 Springer, 2010, pp. 45-65, Springer-Verlag, Berlin, Heidelberg, 2010.

[11] Clinical Decision Support Systems: State of the Art, Agency for Healthcare Research and Quality U.S. Department of Health and Human Services, Prepared by: Eta S. Berner, Ed.D. Department of Health Services Administration University of Alabama at Birmingham AHRQ Publication No. 09-0069-EF June 2009.

[12] C. GAINDRIC, Abordări sistemice în luarea deciziilor, UASM, IMI, Chișinău, 2017, 156 p.

[13] BOIAN, E.; CIUBOTARU, C.; COJOCARU ,S. , MAGARIU, G.; VERLAN, T, ROGOJIN, IU. Sistem de instruire asistată de calculator pentru morfologia limbii române. Lucrările atelierului „Resurse lingvistice şi instrumente pentru prelucrarea limbii Române", Iaşi, noiembrie 2006, pp. 135-139.

[14] BOIAN, E.; CIUBOTARU, C.; COJOCARU ,S.; COLESNICOV, A.; DEMIDOVA, V.; MALAHOV, L.; MAGARIU, G.; VERLAN, T. Tehnologii pentru generarea sistemelor de instruire, dicţionarelor electronice specializate şi ghidurilor lingvistice. Proceedings of the 5th International Conference on „Microelectronics and Computer Science", Volume II, September 19-21, 2007, Chişinău, Moldova, UTM, pp. 20-23.

(Ţiţchiev Inga, Gaindric Constantin) Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chişinău, Moldova

*E-mail address*: inga.titchiev@math.md, constantin.gaindric@math.md