Vladimir Andrunachievici Institute of Mathematics
and Computer Science

Intelligent information systems for solving weakly-structured problems, processing knowledge and big data

Edited by
Svetlana Cojocaru
Constantin Gaindric
Inga Țițchiev
Tatiana Verlan

Chisinau • 2022

CZU 004.9+519.7(082)

I-58

VLADIMIR ANDRUNACHIEVICI INSTITUTE OF MATHEMATICS AND COMPUTER SCIENCE
5, Academiei str., Chisinau, Republic of Moldova, MD 2028
Tel: (373 22) 72-59-82, Fax: (373 22) 73-80-27
E-mail: imam@math.md
WEB address: http://www.math.md

Editors: Prof. Svetlana Cojocaru, Prof. Constantin Gaindric, Dr. Inga Țițchiev, Tatiana Verlan

Authors are fully responsible for the content of their papers.

Descrierea CIP a Camerei Naţionale a Cărţii

Intelligent information systems for solving weakly-structured problems, processing knowledge and big data / edited by Svetlana Cojocaru, Constantin Gaindric, Inga Țițchiev, Tatiana Verlan. – Chişinău : Institute of Mathematics and Computer Science, 2022 (Valinex). – 255 p. : fig., tab.

Antetit.: Vladimir Andrunachievici Inst. of Mathematics and Computer Science. – Referinţe bibliogr. la sfârşitul art. – 150 ex.

ISBN 978-9975-68-462-0.

004.9+519.7(082)

ISBN 978-9975-68-462-0

This issue is supported by the research project 20.80009.5007.22, "Intelligent information systems for solving weakly-structured problems, processing knowledge and big data".

Preface

Svetlana Cojocaru, Constantin Gaindric,
Inga Țițchiev, Tatiana Verlan

The project "Intelligent information systems for solving weakly-structured problems, processing knowledge and big data" was conceived as an attempt to provide scientific assurance of the efforts undertaken in the Republic of Moldova in certain areas of digital evolution, but also with practical recommendations in the development and interoperability of information systems intended for the population. The results obtained along the way were presented at several conferences and workshops related to the project topics. We believe that the overall presentation of the project results will be facilitated by the publication of a separate volume that we present to the interested reader.

The massive use of information and communication technologies revolutionizes the development of modern society and consistently contributes to the implementation of the digital society concept. The Digital Agenda for Europe 2020-2030 focuses on the profound changes brought about by digital technologies, developing and strengthening digital services and markets. Based on two strategic communications, namely "Shaping Europe's Digital Future" and "Europe's Digital Decade", the Commission has set out the specific actions it will take to contribute to the creation of safe and secure digital services and markets, and human-centric and trustworthy artificial intelligence (https://www.europarl.europa.eu/ftu/pdf/en/FTU_2.4.3.pdf).

"The Digital Compass" for this decade consists of four sections, which set out the objectives to be achieved in the following areas:

1. *Skills*, which foresees reaching a capacity of 20 million IT specialists, respecting gender plurality; training of Basic Digital Skills for 80% of the population;

2. *Digital transformation of businesses*, which aims to have 75% of EU companies using cloud technologies, Artificial Intelligence (AI), and Big Data, and 90% of SMEs should achieve at least basic digital intensity. In relation to the notion of "Unicorn Companies", which refers to start-ups that reach a valuation of $1 billion without being listed on a stock exchange, the number is expected to double.

3. *Secure and sustainable digital infrastructures* are the third part of the compass dial, and it sets some benchmarks to be achieved during the decade (Gigabit for everyone, 5G everywhere, the first computer with quantum acceleration, etc.).

4. From the point of view of the topics of our project, the provisions of the Agenda, grouped in the fourth quadrant under the heading *Digitalization of public services*, are of particular importance, as they require the achievement of a 100% rate of development and use of key online services, ensuring a 100% rate of access of citizens to their medical records, as well as 80% use of a digital ID.

The project topic, *Intelligent information systems for solving weakly-structured problems, processing knowledge and big data*, concerning research in the field of intelligent information systems with applications in three social domains: medicine, education, and culture, fits perfectly into this context. In relation to these domains, the project proposes the research and development of information systems aimed at solving scientific and social problems that are, as a rule, the weakly-structured ones, operate with large volumes 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, multi-casualty disaster mitigation management, automation of the process of design and generation of digital content for computer-assisted learning (e-learning). The proposed solutions will take into ac-

count 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 is placed on Romanian works printed in the Cyrillic script, covering the period of the 17th-20th 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.). A universal platform, which prototype is under development in the project, represents a tool for a wide circle of users, offering them assistance at the stages of preprocessing, recognition, alignment, and post-processing of heterogeneous prints. Following the purpose of making information processing more efficient, the solution to problems of design and implementation of distributed computing systems is proposed 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, are proposed and analyzed.

Part 1, "Concepts and tools for interpreting and evaluating information", includes a series of concepts and tools for interpreting and evaluating information dealing with the role of data in contemporary society technologies, generating and visualizing graphical representations of finite automata, collaborative learning modeled by Petri nets, developing augmented artifacts based on the student learning style approach, the use of the concept of crowdsourcing in the discovery of Moldovan cultural heritage through e-Moldova portal, and the examination of the application of artificial intelligence strategies in the Republic of Moldova compared to the countries of the European Union.

Part 2, "Platform for the digitization of heterogeneous documents", is dedicated to the research and development of the platform for the digitization of heterogeneous documents.

It describes the convergent technology in the development of information systems for the recognition of heterogeneous documents and processing the documents with heterogeneous content; the research and exploitation of the Romanian language lexicon in a general Romanian context; and a model of font classification for Romanian Cyrillic documents.

Part 3, "Intelligent information system structures, databases, and knowledge bases for medical triage and diagnostic applications", covers intelligent structures of information systems, databases, and knowledge bases for medical triage and diagnostic applications.

The following aspects are described: formalization of decision-making knowledge and reasoning for victims prioritization; an approach to developing patient diagnosis information structure as a taxonomy in mass casualty disaster management; advanced prehospital triage based on vital signs in mass casualty situations; an artificial intelligence-based response framework for mass casualty management; mass casualty incident triage tools.

Part 4, "Automatic content generation systems for computer-assisted training". Automated content generation systems for computer-assisted learning are needed not only in pandemic cases; they are additional alternative sources including lifelong learning for adults. This theme is presented in Part 4.

Important aspects in evaluating the credibility of unstructured information, e-learning content processing methods and their solutions, the generation and use of educational content in adaptive learning, and a Web crawler model for expanding the initial search domain are presented here.

Part 5, "Systemic concept of the heterogeneous multi-cloud platform and methods of realizing the execution environment of imaging information processing applications", describes how the following problems are examined: development of the computing infrastructure to support Open Science

in Moldova; the actualization of the cloud infrastructure to support research activities, distributed computing infrastructure for the development of complex applications, incident handling, and protection of personal data in medical imaging systems.

The results presented in this volume could form the basis of recommendations for the use of intelligent information systems as development tools in areas that provide:

- the valorization of the national book heritage (including old ones);
- specific medical diagnostic support for doctors and paramedics in management of multiple-casualty disaster mitigation;
- design and generation of digital content for computer-assisted learning for different categories of users.

The demo versions and prototypes of the developed systems (in desktop and network versions), described here, will be proposed for testing and use in the digitization of heterogeneous documents printed in Romanian with Cyrillic characters in the 17th-20th century, patient triage and medical diagnosis, automation of content generation for e-learning platforms, and realization of the execution environment of medical imaging information processing applications.

Svetlana Cojocaru[1], Constantin Gaindric[2],
Inga Țițchiev[3], Tatiana Verlan[4]

Vladimir Andrunachievici Institute of Mathematics and Computer Science
5, Academiei street, Chisinau, Republic of Moldova, MD 2028
E–mails: [1]svetlana.cojocaru@math.md
[2]constantin.gaindric@math.md
[3]inga.titchiev@math.md
[4]tatiana.verlan@math.md

Part 1

Concepts and tools for interpreting and evaluating information

Generation and visualization of graphical representations of finite automata

Constantin Ciubotaru

Abstract

Algorithms are proposed for automatic generation of graphical representations of finite automata (FA) and their visualization. Known methods are modified to explore the specific properties of FA: existence of a single source node, well-defined information flow, absence of isolated nodes. LaTeX/TikZ files are generated for drawing and visualization.

Keywords: finite automata, graph drawing, LaTeX, TikZ.

1 Introduction

Theory and practical application of automata, in particular finite automata (FA), is one of the oldest and most actively studied fields of computer science.

Along with traditional applications related to compiler design, artificial intelligence, word processing, modern applications are currently being developed for natural language processing, speech recognition, software modeling, testing, probabilities (Markov chains), video games, image processing, cryptography etc. The development of the applications inevitably implies the need to carry out equivalent transformations on FA [1], such as:

- elimination of inaccessible and nonproductive states,
- removing ε-transitions,
- converting nondeterministic FA to an equivalent deterministic FA,
- minimizing FA and others.

For most of these transformations it is very useful to represent *FA* as a graph. If analytical and tabular representations do not involve difficulties for visualization, the graphical representation and visualization of *AF* is more difficult.

One graph may be drawn in several ways. Some images may be simple, comprehensible, having an attractive aesthetic appearance, others – more difficult to notice, with a failed structuring. For the finite automaton that recognizes the language $L = \{0,1\}^*\{00,11\}\{0,1\}^*$, there may be drawn several graphical representations. For example, the graph in Figure 1(a) contains multiple intersections of edges, that interfere with visualization and recognition tracking process. Such representations due to the intersections of the edges are called "spaghetti" representations.

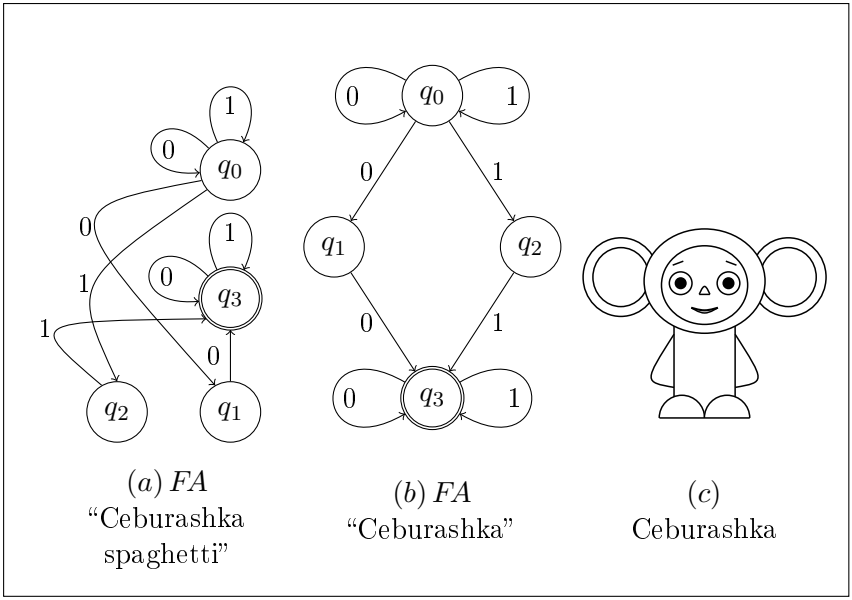


Figure 1. Graphical representation of the "Cheburashka" *FA*

In Figure 1(b) the graph represents the same *FA*, but, unlike the first graph, this one is easier visualizated and obviously illustrates

the process of recognizing language strings. This FA is named as "Cheburashka" in association with the hero of the famous cartoon movie [1] (Figure 1(c)). Such comparation helps to memorize this automaton.

Having several representations of the same graph, it is quite easy to choose the most suitable variant, especially having at its disposal a series of appreciation criteria, usually of an aesthetic nature. For example, the graph must fit into a given and limited space, to contain as few intersections of the edges as possible, to avoid sharp curves, to respect the proportions regarding the length of the edges and the values of the angles of incidence, to favor the elements of symmetry and concentration of nodes, to use suitable shapes for nodes, to respect the orientation of the information flow (from top to bottom or from left to right). For finite automata the general flow of the information will always be oriented from initial state to final states. It is quite difficult to transmit these criteria to the computer. Here intervenes not only the problem of formalizing the criteria, but also the fact that some of them are contradictory. Inevitably there arise compromise situations.

2 Proposed solutions

Starting with the FA definition (or analytical representation), representation of the FA is generated automatically in the form of a graph $\Gamma = (Q, E, F)$, where Q is the set of nodes (states of the FA), E – the set of edges (q_i, a, q_j), and F – the set of final state nodes. In order to obtain the graphical representation, a compiler was developed that generates for this graph an LaTeX/TikZ [2]–[5] file.

The compiler randomly generates and uniformly distributes the node coordinates. This means that between any two nodes there will be respected the minimum distance specified in advance according to the number of nodes.

Figure 2 shows the schema of the developed applications and their interaction.

Generating LaTeX files for analytical and tabular representations is done using `\tabto` and `\tabular` packages and is relatively simple.

[1] `https://en.wikipedia.org/wiki/Cheburashka`

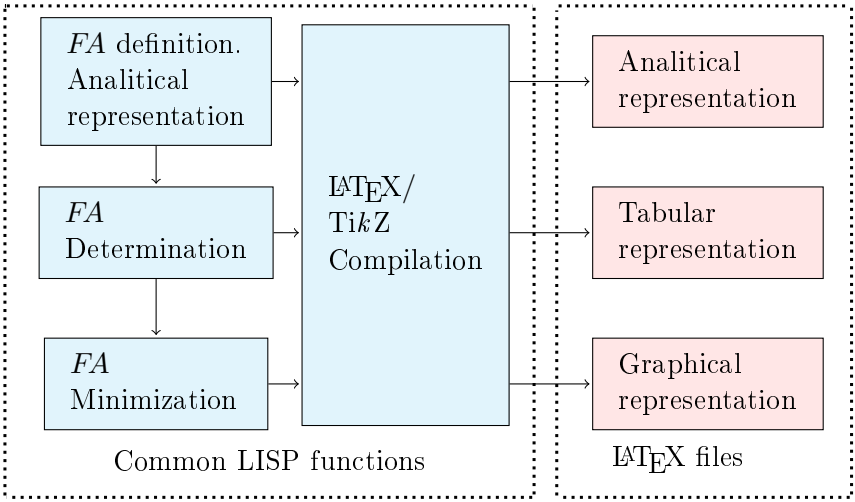


Figure 2. Scheme of the developed applications

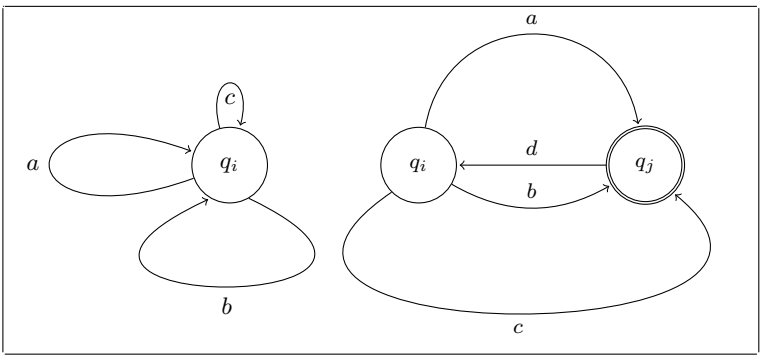


Figure 3. Visualization of the generated fragments

```
\node[state](q_i)at (3.5,3.5){\scriptsize{$q_{i}$}};
\node[state,accepting](q_j)at(6.5,3.5){\scriptsize{$q_{j}$}};
```

a) Templates for nodes.

```
(q_i)edge[to path={(\tikztostart.-60).. controls
    (4.5,1.5)and (-2.6,1.5) .. node[below]
    {\footnotesize $b$} (\tikztotarget.-120)}](q_i)
(q_i)edge[out=200,in=160,looseness=20]node[left]
    {\footnotesize $a$} (q_i)
(q_i)edge[loop above]node[below]{\footnotesize $c$} (q_i);
```

b) Templates for loops.

```
(q_i) edge [out=80,in=100,looseness=1.5] node [above]
 {\scriptsize $a$} (q_j)(q_i) edge [bend right=30] node
 [above]{\scriptsize $b$} (q_j)
(q_i) edge [to path={(\tikztostart.-135).. controls (0.0,1.0)
  and (9.5,1.0) .. node [below]{\scriptsize $c$}
 (\tikztotarget.-45)}](q_j)
(q_j) edge node [above]{\scriptsize $d$} (q_i);
```

c) Templates for edges.

Figure 4. Generated templates

Several difficulties arise during generating graphical representations. It conserns the way the nodes are placed, the loops and edges are drawn, and the tendency to obtain an aesthetic and comprehensible drawing. Full automation of this process is virtually impossible, because the formulated criteria are difficult to formalize, and sometimes they are even contradictory [6]. The solutions, proposed in this paper, will help the user to build an acceptable graphic structure.

The developed compiler exploits, along with the possibilities of the LaTeX system, the package `\tikz` and its libraries, especially the libraries `automata, positioning, arrows`. It should be mentioned that there exists the possibility to highlight the final and the initial states and rich arsenal for drawing loops/edges.

Several possible variants are inserted in Figure 3. For each edge (loop) functional templates are generated, one is active, the others are commented and offered to the user in case he wants to intervene manually.

These templates are inserted in Figure 4.

3 Example

An *FA* was selected as an example for which there were automatically generated the deterministic and minimized equivalent models. LaTeX files for tabular and graphical representations were generated for all these constructs. All these are inserted in Figures 5, 6.

4 Conclusion

The applications proposed in the paper are used in the training process for studying finite automata. Of course, manual intervention is still required, but the functional version LaTeX, obtained as a result of the compilation, is very useful. Upcoming the involvement of the Sugiyama framework is envisaged, which broadly minimizes the number of the edges intersections.

EXAMPLE

Analitical definition

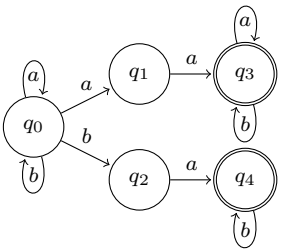
$AF=(Q,\Sigma,\delta,q_0,F)$.
$Q=\{q_0,q_1,q_2,q_3,q_4\}$,
$\Sigma=\{a,b\}$,
$F=\{q_3,q_4\}$
$\delta(q_0,a)=\{q_0,q_1\}$,
$\delta(q_0,b)=\{q_0,q_2\}$,
$\delta(q_1,a)=\{q_3\}$,
$\delta(q_2,a)=\{q_4\}$,
$\delta(q_3,a)=\{q_3\}$,
$\delta(q_3,b)=\{q_3\}$,
$\delta(q_4,b)=\{q_4\}$

Tabular reprezantation

$AF=(Q,\Sigma,\delta,q_0,F)$.
$Q=\{q_0,q_1,q_2,q_3,q_4\}$,
$\Sigma=\{a,b\}$,
$F=\{q_3,q_4\}$

q	$\delta(q,a)$	$\delta(q,b)$
q_0	$\{q_0,q_1\}$	$\{q_0,q_2\}$
q_1	$\{q_3\}$	$\{\}$
q_2	$\{q_4\}$	$\{\}$
q_3	$\{q_3\}$	$\{q_3\}$
q_4	$\{\}$	$\{q_4\}$

Graphical reprezentation



Deterministic FA, generalized states

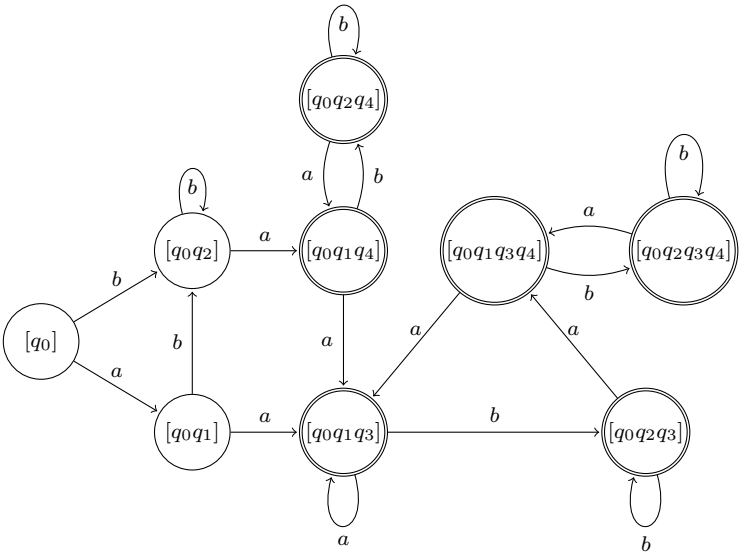


Figure 5. Example. Page 1

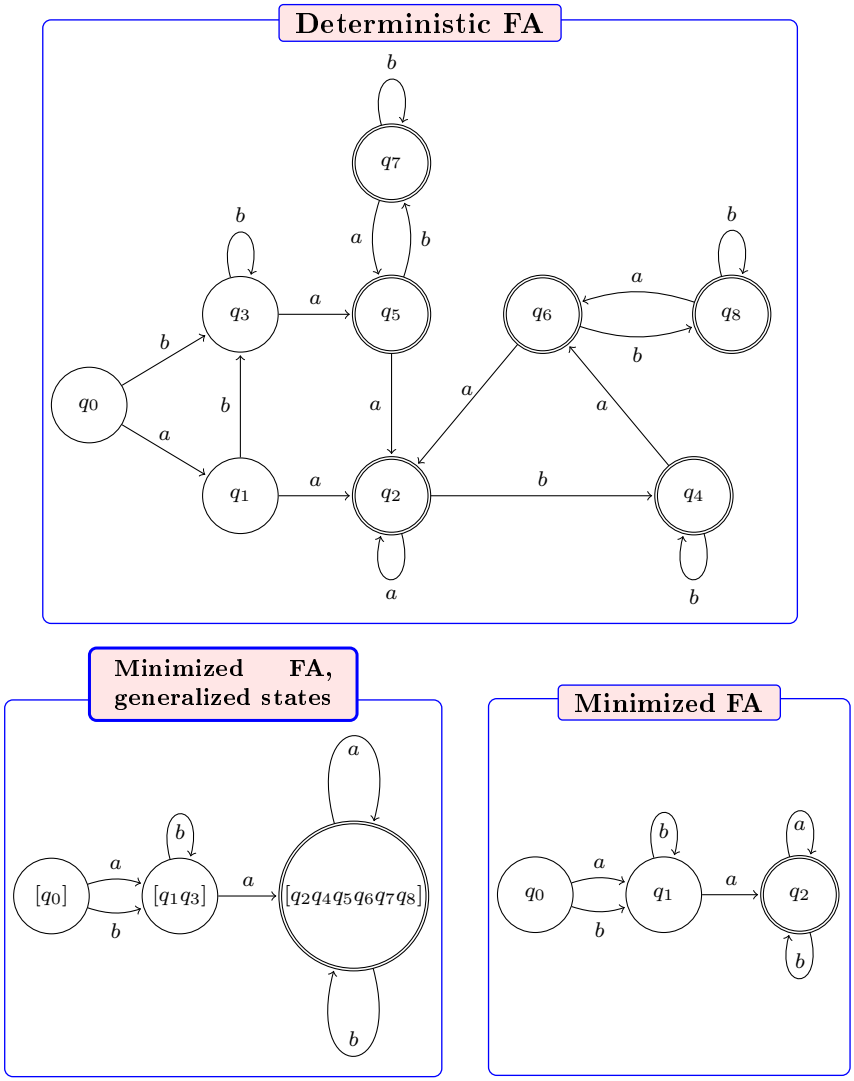


Figure 6. Example. Page 2

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 “Intelligent information systems for

solving ill-structured problems, processing knowledge and big data".

References

[1] J.E. Hopcroft, R. Motwani, and J.D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison Wesley, 1979, 427 p.

[2] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion.* Addison-Wesley, Reading, Massachusetts, 1993.

[3] Stefan Kottwitz. *LaTeX Cookbook.* Packt Publishing, 2015, 387p.

[4] *The TikZ and PGF Packages. Manual for version 3.1.6.* Institut für Theoretische Informatik Universität zu Lübeck, 2020, 1320p.

[5] Till Tantau. *Graph Drawing in TikZ.* Journal of Graph Algorithms and Applications, 17 (4), 2013, pp.495–513.

[6] G. Di Battista, P. Eades, R. Tamassia, and I.G. Tollis. *Graph Drawing. Algoriithms for the visualization of graphs.* Prentice Hall, 1999, 397 p.

[7] K. Sugiyama, S. Tagawa, and M. Toda. *Methods for visual understanding of hierarchical system structures.* IEEE Trans. Syst. Man. Cybern., 11(2), 1981, pp.109–125.

[8] N.S. Nikolov. *Sugiyama Algorithm.* Encyclopedia of Algorithms - 2016 Edition, Springer 2016, pp. 2162–2166.

Constantin Ciubotaru

Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: chebotar@gmail.com

Parallel architecture and software by workflow Petri nets

Inga Titchiev

Abstract

In this article we illustrate the possibilities offered by Petri nets in modeling the issues related to distributed data processing system.

Keywords: Petri nets, workflow Petri nets, parallel computing, High performance computing.

1 Introduction

High performance computing (HPC) infrastructure development offers many advantages for solving problems in various fields (bioinformatics, physics, chemistry, mathematical modeling, web servers and databases, optimization of business decisions and medicine). Development of HPC is characterized not only by increasing the number of elements involved in data processing, but by presenting the relationship between them and the management of interactions by very complex structure. Such interactions contributed to the occurrence of new problems related to the analysis, modeling and representation of causal relations in complex systems with many objects acting in parallel. In order to integrate the physical (architecture) and logical (software) components of a distributed data processing system, a method like Petri nets will be used.

2 Software by workflow Petri nets formalism

Petri nets formalism, during over the five decades since they were introduced, showed great flexibility in dealing with many types of practical problems and a great capacity for expansion by incorporating some

more complex points of view. Modeling of software component, represented by block scheme, as usually, is performed at the component level. Transitions are associated with program actions: calculations and decisions. For the interpretation of a Petri net, an interpretation must be made for each transition. The translation of the elements of block scheme through workflow Petri nets is presented in Figure 1.

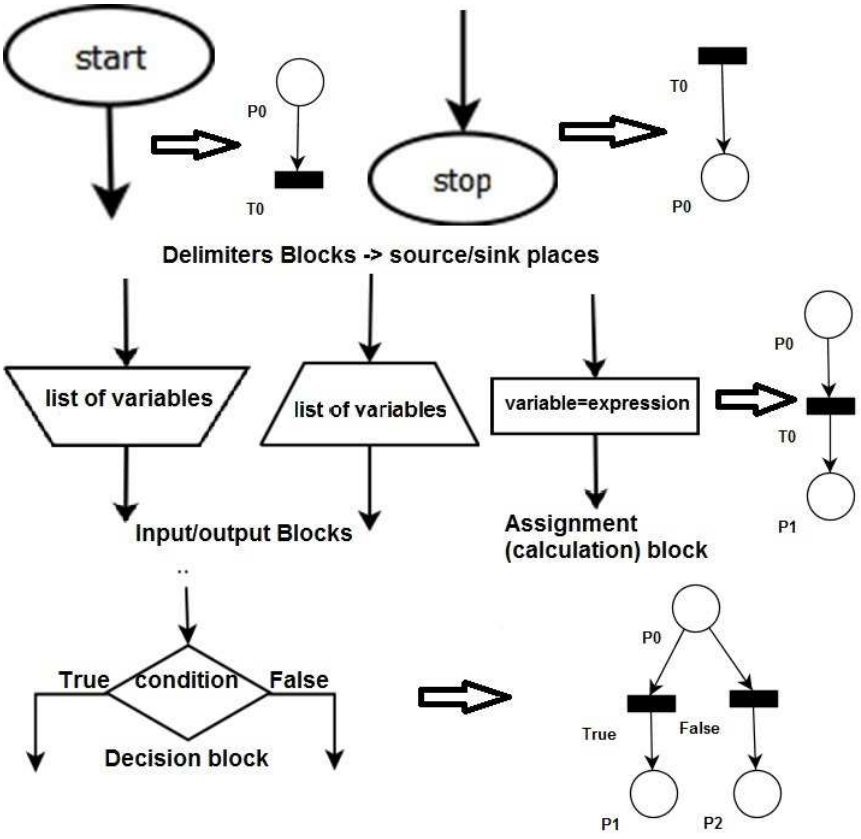


Figure 1. Block scheme elements by Petri nets

The transitions for the calculation actions have a single input and a single output, there can be no conflict for a transition representing a calculation, since its input location is no longer the input location for

any other transition. Decision-making can bring conflicts into the net, but in a very narrow way: any choice can be made. The choice can be made either in non-deterministic way (i.e. randomly), or it can be controlled by the same external force (i.e. an agent) that calculates the truth or falsity of the decision and forces the triggering of the correct transition.

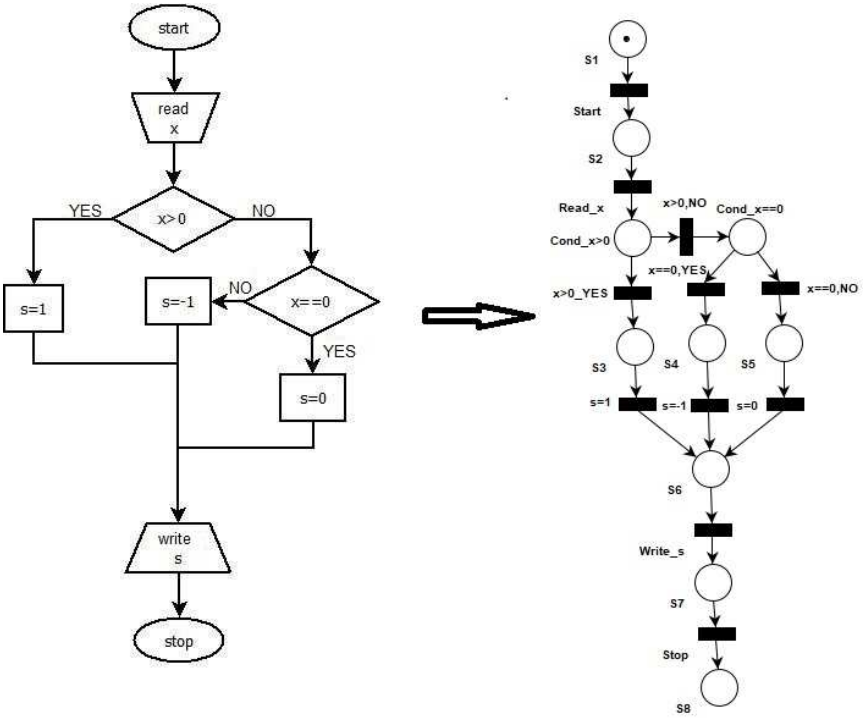


Figure 2. Block scheme and workflow Petri net

3 Parallel architecture

Ability to shape overlap and easily to combine subsystems using Petri nets, makes them useful for modeling of the more complex hardware components. Computational systems consist of many components, and

many designers try to increase speed through parallel execution of certain functions. This is why Petri nets are a very suitable representation for such type of system. In order to convert application (or part of the application) from Sequential computing mode (when application runs on one computer and its run time depends on the computing capacity of the computer) to the parallel one, it is necessary to adopt one of the following parallelization technologies:

1. Parallel computing on one multicore computer – OpenMP technology.

2. Parallel computing on several computers, using one core on each (MPI Message Passing Interface technology).

3. Hybrid technology – using both OpenMP and MPI.

4. Distributed or Grid-computing – the ability to simultaneously run the same serial applications on multiple computers – distributed computing (computers of various capacities are united in a parallel computing system by the local and global networks).

Parallelism (competition) can be introduced in several ways. We consider the case of two concurrent processes. Each process can be represented by a workflow Petri net. Therefore, the composed Petri net, which is simply the union of Petri nets for each of the two processes, may represent concurrent execution of the two processes. Initial marking of compound workflow Petri net has two tokens, with one in each source place representing the beginning of each process. This fact introduces parallelism that can not be represented by a logical scheme and for which such representation as a Petri net is a very useful solution.

Another approach is to consider how parallelism can be introduced in a normal process in a computer system. Branching operations (FORK) and union (JOIN) ones, discussed in [1], [4], are considered.

Parallelism is useful in solving a problem if concurrent processes can cooperate in solving the problem. Such cooperation involves common processes information and resources. Shared access to information and resources must be controlled in order to ensure correct functioning of the system.

The Vladimir Andrunachievici Institute of Mathematics and Computer Science's cluster allows running Sequential, Parallel(Open MP, MPI and Hybrid) applications and Grid cluster.

A variety of coordination problems has been proposed in the literature to illustrate the types of problems [2], [3], [5] that may arise between cooperating processes.

4 Conclusion

In this article it was shown that Petri nets is a convenient formal method for modeling information flow represented by block scheme. Thus, it was shown how Petri nets have been used for representation of parallel processes in order to better understanding of these processes.

Acknowledgments. The research project 20.80009.5007.22, "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" has supported part of the research for this paper.

References

[1] J.B. Dennis and E.C. Van Horn. *Programming Semantics for Multiprogrammed Computations*. Communications of the ACM, 9(3), pp. 143–155, March 1966.

[2] E. W. Dijkstra. *Solution of a problem in concurrent programming control.* Communications of the ACM, vol. 8(9), pp. 569, 1965.

[3] M. Raynal. *Algorithms for mutual exclusion.* The MIT Press, 1986. Translation of: Algorithmique du parallelisme, 1984.

[4] I. Titchiev, N.Iliuha, and M.Petic. *Workflow Petri nets used in modeling of Parallel architectures*. Proceedings of the International Conference on Intelligent Information Systems, August 20-23, Chisinau, Republic of Moldova, 2013, pp. 163–167.

[5] Syed Nasir Mehmood, Nazleeni Haron, Vaqar Akhtar, and Younus Javed. *Implementation and Experimentation of Producer-*

Consumer Synchronization Problem, International Journal of Computer Applications (0975 – 8887), vol. 14, no.3, January 2011, pp. 32–37.

Inga Titchiev[1, 2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
[2] Tiraspol State University

E–mail: inga.titchiev@math.md

Discover the Moldovan Cultural Heritage through e-Moldova Portal by Using Crowdsourcing Concept

Olesea Caftanatov, Tudor Bumbu, Lucia Erhan, Iulian Cernei, Veronica Iamandi, Vasile Lupan, Daniela Caganovschi, Mihail Curmei

Abstract

Our research aims to analyze the needs of the people and, by using modern technologies, to develop a portal that will use the wisdom of crowds to contribute to the development of the Republic of Moldova. In this paper, we present the first version of e-Moldova portal.

Keywords: crowdsourcing, portlets, digital cultural heritage, Flarum core.

1. Background research

Technology is present more than ever in our lives and it is difficult to imagine our daily routine without it. Over the years, technology has paved the way for multi-functional devices. As a result, it made our lives easier, faster, better, and more fun. The Republic of Moldova still has work to do in terms of implementing modern technologies in everyday life, but it has a great potential to manifest itself in the IT field in the near future. Our team pursues the following objectives:

- researching the needs of people and bringing new technologies to make our country a prosperous one in the near future;
- searching fields in which our expertise will be more useful for our country;

- collecting, creating, and managing informational resources regarding our country, also providing free access to them;
- digitization of our cultural heritage by involving crowds wisdom in our project;
- developing tools that will help the mass to rediscover our country.

In this paper, we will present the results of one year of our work in pursuit of these objectives and describing the first version of our project.

2. E-Moldova Portal

Our team also brings its contribution to the promotion and development of the Republic of Moldova through the development of the e-Moldova portal [1], a platform dedicated to digitization of our cultural heritage and complementing it with new knowledge by involving the wisdom of crowds in exploring "portlets" on history, culture, geography, demography, politics, economics, career, education, wellness, press, informational technologies, and, of course, the digital thesaurus (see Fig.1).

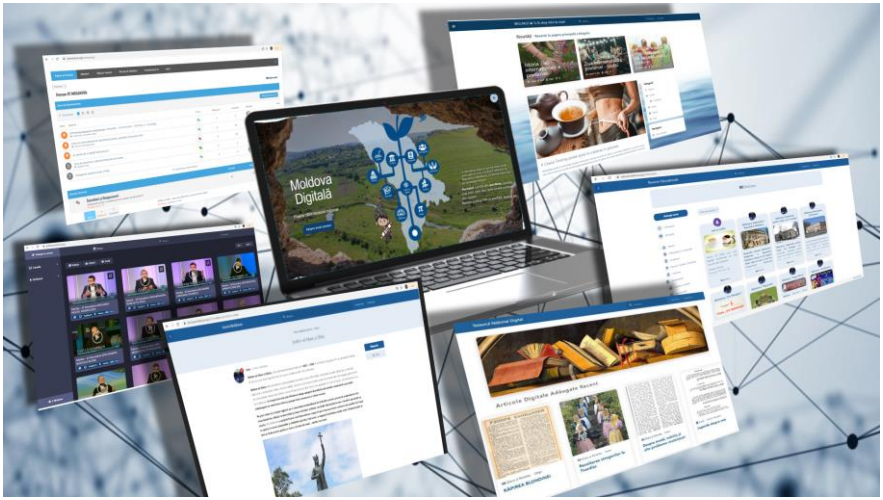


Figure 1. E-Moldova portal main interface with six portlets

According to the Lexico dictionary powered by OXFORD [2], the term *"portlet"* in the informational field means *"an application used by a portal website to receive requests from clients and return information"*.

Our team essentially sees a portlet as a specialized content area within a web page that occupies a small place and serves as a hotspot leading towards a large body of information (the word ***portlet*** has the same meaning as the word ***door***). The nature of information can variate from typical posts, video lessons, galleries, and access to other resources or even tools. Nevertheless, all portlets are designed to use the wisdom of crowds in order to further evolve. In other words, our project is based on the concept of "*crowdsourcing*" developed by James Surowiecki [3].

The meaning of this concept can be conveyed by the formula: if two heads are better than one, a hundred heads will be great. We constantly encounter applications based on this approach. An eloquent example of the use of mass intelligence is Wikipedia.

Thousands of Wikipedia users have created an encyclopedia that, according to [4, 5], is considered to be a good source of almost the same accurate information as Britannica. One of our objectives is to create a tool that will enable crowds' intelligence to create a body of knowledge about the Republic of Moldova presented in a digital format.

To this day, we have initialized 12 portlets that can be accessed through the e-Moldova portal (see Fig. 2). The portal's interface was intended to be as simple as possible: a menu is displayed as a grape, and each berry represents one of the portlets.

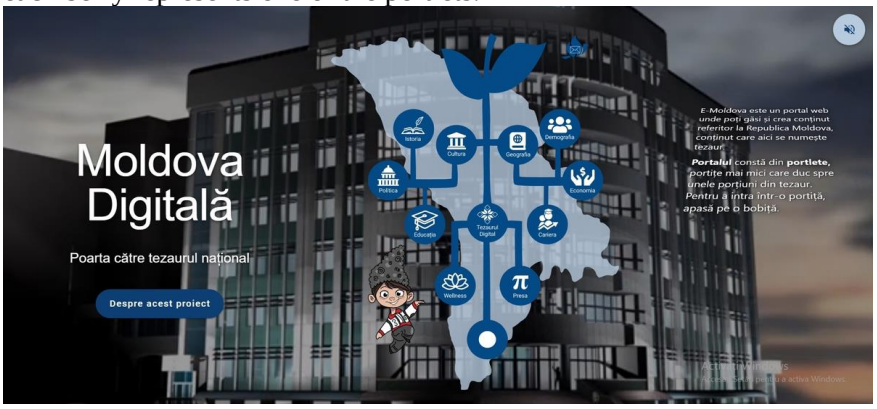


Figure 2. The interface of the e-Moldova portal

Each berry has a toolkit with a short description of portlets. In the background, we play a video with places that can be seen in our country or events that took place in it. Thus, the background is a component of the user interface that adds to the knowledge about Moldova.

The portal's layout consists of 3 sections. On the left side, we have the title and a button that redirects to a discussion board, where anyone can leave some feedback for improvement of the portal or even its portlets. On the right side of the layout, we displayed an animated text with some tips on how to access our portlets. In the upper right corner, we added a voice button that recites the text in 3 languages (Ro, Ru, and Eng). The main part of the layout is the center, where our menu is displayed as a grape on the outlined map of our country. In addition, near the right leaf, we have a pigeon with the envelope, for the case when the accessed users wish to leave a message to the administrator.

Another digital element is in the process of developing, it is about the Guguța avatar, and this animated character will navigate through our interface and help users to better understand our project.

3. Tools for Portlet development

To develop our portlets, we used Flarum discussion platform [6] as a core and Wordpress content management system [7]. Flarum is built with Laravel framework [8] so it's quick and easy to deploy. Moreover, its interfaces are powered by Mithril, a performant JavaScript framework for single-page applications. Flarum's architecture is flexible, and it is equipped with a powerful Extension API. We easily can extend, customize or even integrate our extensions to the base platform. All portlets, except IT forum portlets, were developed using Flarum Core. Regarding the IT forum portlet, we used the WordPress platform. We also intend to use a wiki engine [9] to bring the possibility for users to collaboratively edit the content (the articles).

4. Portlets

As we have mentioned before, we have initialized 12 portlets that can be accessed through the e-Moldova portal, but in this paper, we described only five of our portlets: *IT Forum, Education, Press, Digital Thesaurus, and Wellness*. We chose to describe these five portlets because they have

more advanced interfaces and functionalities. The other initialized portlets have the basic functionality for content creation and management. Going forward, we intend to add specific functionality for each one of them.

4.1 IT Forum Portlet

One of the first portlets initialized was the IT forum. This portlet brings together a wide range of current information technologies, including useful information for pupils, students who are interested in learning a programming language or learning more about some branches of IT (see Fig. 3).

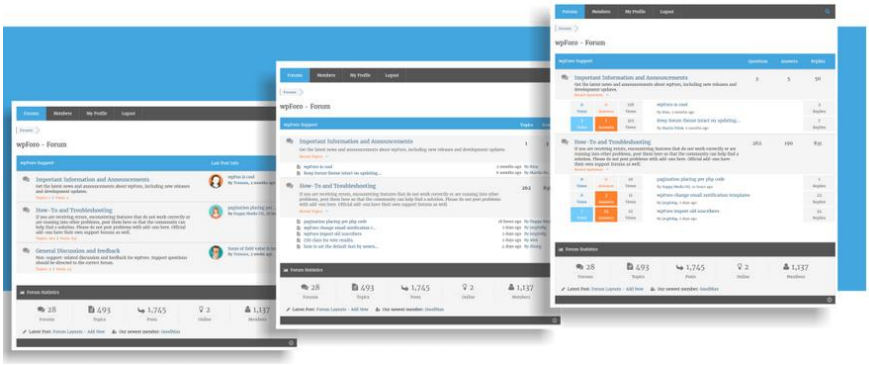


Figure 3. IT Forum portlet homepage interfaces

As a core for IT Forum, we use wpForo [10] plugin from WordPress platform made by gVectors Team. wpForo has become a new generation of Forum Software. It has multi-layouts such us: The “Extended”, “Simplified” and “Question & Answer”. Layouts fit almost all types of discussions needs.

Beautiful, modern, and informative profile system, with member pages as a statistic, bio, settings, activity, and subscriptions. It has a user rating system based on several posts. Nice Badges and Member Rating Titles per reputation level. Powerful moderation tool, fully customizable.

Our team for one year created 580 topics with 665 posts organized in 47 under IT forums. Each post that was created was checked on plagiarism by using PREPOSTSEO Tool [11]. Only content that has an average of 70% unique was approved and published. Thus, through passion, we develop a community for IT geeks.

4.2 Education Portlet

The Education portlet (see Fig. 4) is a collection of information resources, intended for four categories of users, grouped in distinct communities such as the community of teachers, parents, pupils, and the community of students. Resources are selected for user education or as an aid in choosing a school or university; courses, sports clubs; the choice of useful tools in education, etc.

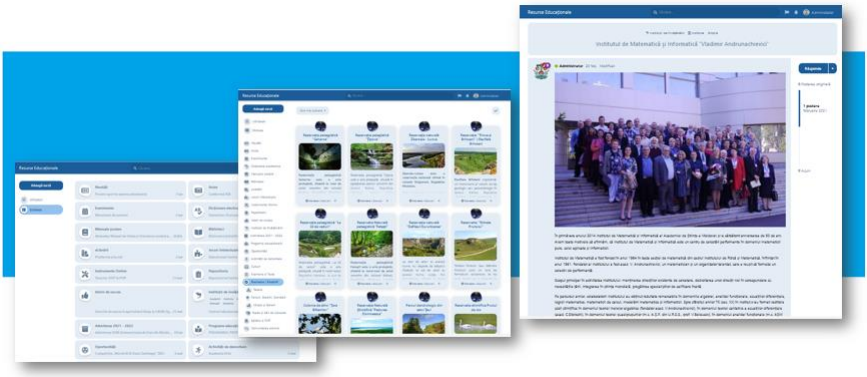


Figure 4. Educational portlet interfaces

Now, the Educational portlet has over 700 information resources presented in the form of cards, which continue to grow every day. Guests have access to sources and even the ability to post a source that may be useful to other users, or may leave feedback in the form of comments. For the Educational portlet, we allow any type of added resources, the minimal requirements are:

- useful resource;
- no uncensured content is allowed.

4.3 Press Portlet

This portlet's mission is to increase the impact of the independent press in the Republic of Moldova and to contribute to the creation and consolidation of the open society (see Fig. 5).

We believe that anyone who has the ability can create an article even if he/she is not a journalist by profession. At some point, every person can contribute with any kind of information.

Our team can moderate the posts created by the masses and generate from the top rating posts a digital newspaper. However, because of little human resources, we started with a compilation of the best TV Shows on three popular Channels: “Publika TV”, “TVC21”, and “10TV”. All the collected resources can be analyzed and get a place in the rating system. In such a way the more useful TV Shows will be posted on top of the list.

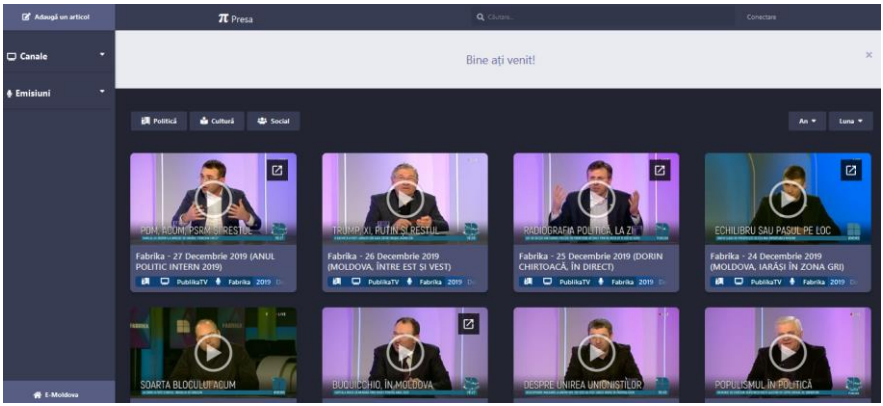


Figure 5. Press portlet homepage interface

4.4 Digital Thesaurus Portlet

The Digital National Thesaurus portlet aims to digitize articles that are part of the Moldovan cultural heritage, using sophisticated information technologies to help you study and learn about the country's past and ancestors (see Fig. 6).

We aim to provide access to articles from the following collections: newspapers and magazines, archival documents, books and manuscripts, collections gathered and selected from past centuries. The word "digitization" refers to a technological process, where an item from some collections is converted and transformed into a digital item, which can then be placed on the Internet.

Such an article can be one or more pages from a newspaper, a magazine or a book, a historical document, but maybe even a photo from a Moldovan village, and so on.

The portal will ensure access to newspapers and magazines printed in Moldovan Cyrillic alphabet. We consider newspapers and magazines as a

priority collection because they are about the lives of Moldovans with "many truths left in the past".

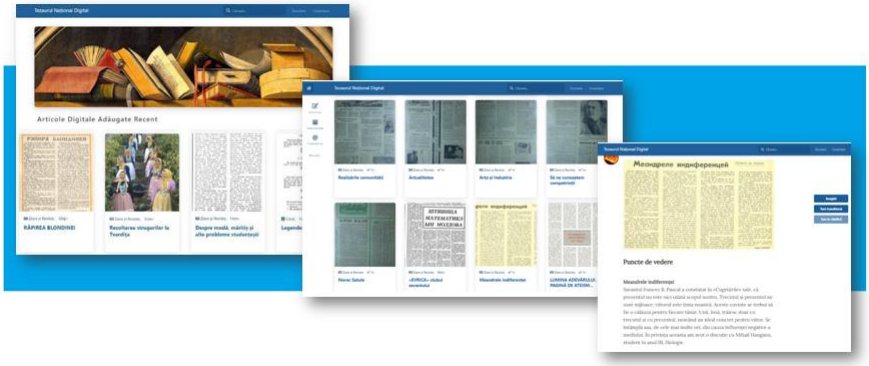


Figure 6. Digital Thesaurus portlet interfaces

Newspapers and magazines from the 20th century are kept in many libraries in the country, and among them, there are the following: the National Library of Moldova, the National Archive of the Republic of Moldova, the Central Scientific Library "Andrei Lupan", the USM Library, etc. The result of this project is the national digital hoard consisting of digital articles.

This portlet differs from the other portlets because it includes technology for digitization. The technology consists of a tool pack for image preprocessing, layout analysis and segmentation, optical character recognition, and transliteration from the Moldovan Cyrillic alphabet to Latin.

4.5 Wellness Portlet

The Wellness portlet was created to inspire people to choose a healthier lifestyle. Through this portlet, our team distributes some results from studies done on various dimensions of well-being what we call Wellness (see Fig. 7).

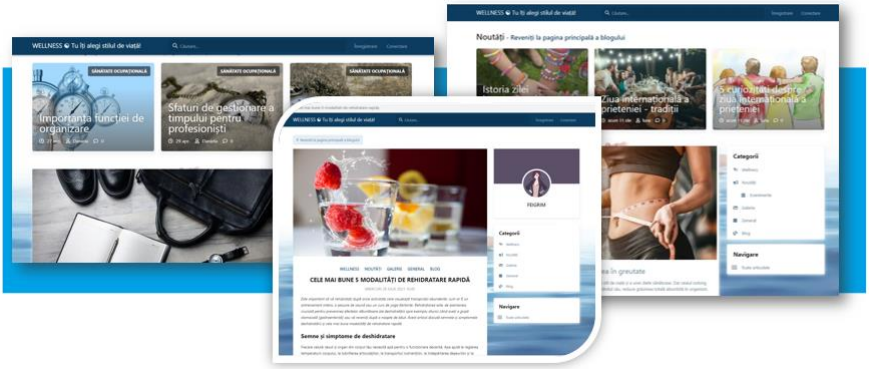


Figure 7. Wellness portlet interfaces

Because wellness is more than getting rid of an illness, wellness is a permanent dynamic process, to become aware, responsible, and to make decisions that contribute to well-being. It refers to change and growth, through which the physical, intellectual, emotional, social, spiritual, occupational levels, as well as the level of the environment, develop. Every dimension is equally vital in the pursuit of optimal health. It is the perfect balance between mind, body, and spirit.

In total, we have created over 230 topics for discussion that can be commented on, appreciated, and distributed on various social networks. The portlet provides free access to nutritional information, recipes, and more. All posts were checked by a plagiarism checker, those that have plagiarized content more than 30%, we return to the author and ask them to redo. Additionally, if there are some scientific articles in other language but are very interesting, we translate them and post. Moreover, at the end of the post, we write that it was a translated work and we share the link to the original article.

5. Conclusion

In this paper, we presented the e-Moldova project that uses the crowdsourcing concept and modern technologies to develop a portal with its portlets. Each portlet leads to a specific field about the Republic of Moldova, be it regarding the economical field or political, etc. In this article, we described the purpose of the main portal and shared some ideas

about the first five portlets. Currently, the content is being created and managed by our team. It seems to be a normal thing in the initial phase. However, we intend to delegate this work to the users and in this manner implement the crowdsourcing approach.

Acknowledgments. We would like to express our sincere gratitude to Dr. Ioachim Drugus, the owner and Dev Lead of e-Moldova portal © (also referenced as e-Moldova ©, or Digital Moldova ©), for continuous support and for giving us the opportunity to study, research and participate in the development of the portal. We hope that those who learned about this project and were inspired by our activity will try to get involved in any form – through feedback, by sending an article for publication, or by suggesting a useful resource. On this occasion, we would like to also express our gratitude to the people on the project "Intelligent information systems for solving ill-structured problems, processing knowledge and big data", who support us, the young researchers, engineers, inventors, that still cannot produce something to be sold.

References

[1] *E-Moldova Portal*, official website: https://emoldova.org/.

[2] *Lexico Dictionary*, website link: https://www.lexico.com/definition/portlet.

[3] J.Surowiecki. *The Wisdom of Crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Doubleday, 2004, pp. 296.

[4] D. Terdiman. *Study: Wikipedia as accurate as Britannica.* Digital post on Cnet website. December 16, 2005. Accessed on August 1, 2021. https://www.cnet.com/news/study-wikipedia-as-accurate-as-britannica/.

[5] N. Wolchover. *How accurate is Wikipedia. Digital post on Live Science Website*. January 24, 2011. Accessed on July 15, 2021. https://www.livescience.com/32950-how-accurate-is-wikipedia.html.

[6] *Flarum*, official website. https://flarum.org/.

[7] *WordPress*, official website: https://wordpress.com/.

[8] *Laravel*, official website: https://laravel.com/docs/8.x.

[9] *Wiki Software*. Wikipedia, https://en.wikipedia.org/wiki/Wiki_software.

[10] *wpForo Plugin on WordPress*, official website: https://wpforo.com/.

[11] *Plagiarism checker Tool*, official website: https://www.prepostseo.com/plagiarism-checker.

Olesea Caftanatov[1], Tudor Bumbu[2], Lucia Erhan[3], Iulian Cernei[4], Veronica Iamandi[5], Vasile Lupan[6], Daniela Caganovschi[7], Mihail Curmei[8]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science; SRL EST Computer
E-mail: olesea.caftanatov@math.md

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science; SRL EST Computer
E-mail: tudor.bumbu@math.md

[3]EST Computer SRL
E-mail: ladyblackserinity@gmail.com

[4]EST Computer SRL
E-mail: iuliancernei@gmail.com

[5]Vladimir Andrunachievici Institute of Mathematics and Computer Science; SRL EST Computer
E-mail: veronica.gisca@gmail.com

[6]EST Computer SRL
E-mail: lupan.01@mail.ru

[7]State University of Moldova; SRL EST Computer
E-mail: cda3721@gmail.com

[8]EST Computer SRL
E-mail: mcurmei@mail.ru

Collaborative learning modelled by High-Level Petri nets

Inga Titchiev

Abstract

The rapid development of internet technology induces a new style of learning, which is different from the traditional one and comes to complete it with new opportunities. The study described in this article the researched management of the learning progress and collaborative issues in distance learning by means of Petri nets.

Keywords: e-learning, collaborative learning, Petri nets, efficiency of learning.

1 Introduction

Increasing the efficiency of the educational system, extension and diversifying the educational offers, continuous training by exploitation the opportunities offered by information and communication technologies are the development priorities of the educational system in the Republic of Moldova.

Until recently, in many countries, distance learning has not been widely used for several objective reasons – mainly due to the insufficient development of technical means of training. Currently, the technical premises for the widespread use of distance learning in education have been created, and the COVID-19 pandemic has energized and accentuated their urgent need.

It is essential to identify the needs of users and to integrate into the system the functionalities that allow them to be satisfied. Even if the emphasis shifts from teacher to student, it is still necessary to obtain information about student progress and the level of collaboration

with other students, so it is proposed to use Petri nets [4] as a tool in modeling the management of these processes.

2 Distance education

Distance education [6] gives to learner flexibility in time and location. Thus, being geographically dispersed, they have opportunities to collaborate and to develop even in crisis situations.

Definition 1 [7]. *Collaborative learning is the educational approach of using groups to enhance learning through working together. Groups of two or more learners work together to solve problems, complete tasks, or learn new concepts.*

Collaborative learning involves new opportunities for learners and new approaches for teachers, their role being no less necessary, as a mentor, thereby optimizing the teaching process by distributing the resources of the trained and organizing activities through new technologies.

In order to increase the efficiency of group learning, information and communication technologies in education are coming. An important role in increasing the effectiveness and efficiency of collaborative learning [5] is the motivation of each member of the group, the number of members, their skills.

The proposed approach of modeling the individual and group route management through Petri nets, the MAETIC learning method (from the French: Pedagogical Method with ICT tools), will be applied, which is based on project-based development.

3 High-Level Petri Nets

High Level Petri-nets (HLPNs) asset:

1. High Level Petri-nets have an intuitive graphical representation and a well-defined semantics that unambiguously define the behaviour of each HLPNs.

2. They are very general and can be used to describe a large variety of different systems.

3. HLPNs have a very few, but powerful, primitives, an explicit description of both states and actions.

4. Are stable towards minor changes of the modelled system.

5. A formal analysis methods allow proving the properties of HLPNs.

6. The two most important analysis methods are known as occurrence graphs and place invariants. Computer tools [8] supporting their drawing, simulation, and formal analysis, exist.

Definition 2 [2]. *A* High-level Petri Nets *is a structure* $HLPN = (P; T; D; Type; Pre; Post; M_0)$*, where*

- P is a finite set of elements called Places.
- T is a finite set of elements called Transitions disjoint from $P(P \bigcap T = \emptyset;)$.
- D is a non-empty finite set of non-empty domains, where each element of D is called a *type*.
- $Type : P \bigcup T \rightarrow D$ is a function used to assign types to places and to determine transition modes.
- $Pre; Post : TRANS \rightarrow \mu PLACE$ are the pre and post mappings with

  $TRANS = \{(t; m) | t \in T; m \in Type(t)\};$

  $PLACE = \{(p; g) | p \in P; g \in Type(p)\}.$
- $M_0 \in \mu PLACE$ is a multiset called the initial marking of the net.

A *Marking* of the HLPN is a multiset, $M \in \mu PLACE$.

A transition is enabled with respect to a *net marking* or in a particular *transition mode*. A transition mode is an assignment of values to the transition's variables, that satisfies the transition condition (i.e., the transition condition is true). The transition's variables are all those

variables that occur in the expressions associated with the transition. These are the transition condition and the annotations of arcs involving the transition.

A finite multiset of transition modes, $T \in \mu TRANS$, is enabled at a marking M iff $Pre(T_\mu) \leq M$.

A step may occur resulting in a new marking M' given by $M' = M - Pre(T_\mu) + Post(T_\mu)$.

3.1 Mapping High-Level Petri Nets for collaborative learning

In this section, we use HLPNs [3] to construct various sequence control in distance learning. Depending on the behavior of the learner we can have different learning paths. In order to identify these paths, we will specify several control sequences that may occur. Based on the same course content, we can have different instructional strategies: linear, choice, and arbitrary traces (which combines the first two).

For linear learning path, the learner progress is in a pre-determined order (Figure 1). In Figure 1, the learner can go to the second topic only after finishing the first one, and so on.

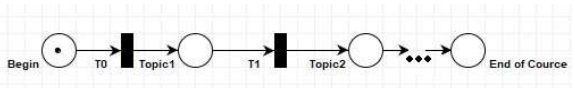


Figure 1. Linear learning path by Petri nets

Linear choice path allows jumping and selecting the next content in the arbitrary order (Figure 2). The learner in Figure 2, can access any topic he wants from n existing ones arbitrarily.

The collaborative learning is an effective method in distance education, thus we propose to model this process for better understanding.

The goal of collaborative learning is to form a group with heterogeneity even if they have different backgrounds, various learning paths and diverse instruction styles. In order to achieve this goal, it is necessary that each learner has the opportunity to make a break point (jump) to obtain additional information from the outside (a sub-net), so that on return he/she can ensure the homogeneity of the group.

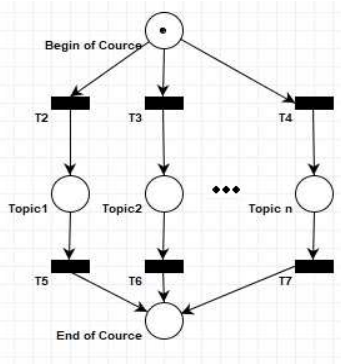


Figure 2. Choice learning path by Petri nets

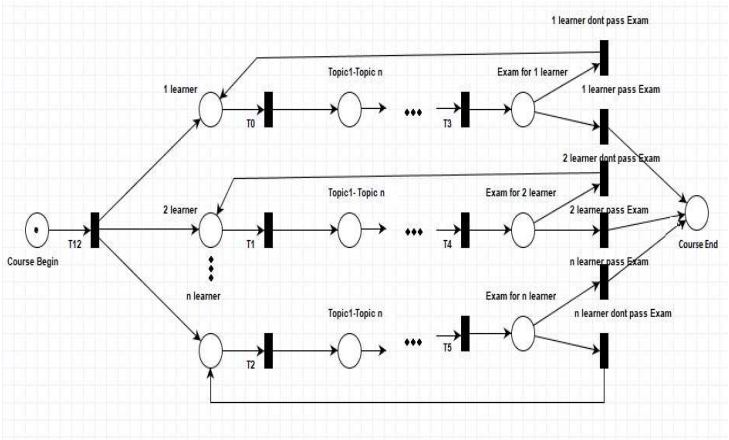


Figure 3. Collaboration modeled by H-L Petri nets

After the modeling of collaboration learning trace, by analysing the HPNs, we can estimate the block, deadlock of the system, learning path in order to improve the process.

4 Conclusion

In this article it was shown that Petri nets is a convenient formal method for modeling the management of the learning progress and collaborative issues. Thus, it was shown how to identify the needs of users and integrate the functionalities of the system for better understanding of these processes.

Acknowledgments. 20.80009.5007.22, "Intelligent Information systems for solving ill-structured problems, knowledge and Big Data processing" project has supported part of the research for this paper.

References

[1] X. He, and T. Murata. *High-Level Petri Nets – Extensions, Analysis, and Applications*. Electrical Engineering Handbook (ed. Wai-Kai Chen), Elsevier Academic Press, 2005, pp.459–476.

[2] K. Jensen. *An Introduction to High-level Petri Nets*. In: Proceedings of the 1985 International Symposium on Circuits and Systems: Kyoto 85, pp 723–726, Kyoto, Japan, 1985.

[3] K. Jensen, and G. Rozenberg. *High-level Petri Nets: Theory and Applications*. Springer-Verlag, 724 p., London, UK, 1991.

[4] S. Cojocaru, M. Petic, and I. Titchiev. *Adapting Tools for Text Monitoring and for Scenario Analysis Related to the Field of Social Disasters*. In: Proceedings of The 18th International Conference on Computer Science and Electrical Engineering (ICCSEE 2016), October 6-7, 2016, Prague, Czech Republic, pp. 886–892.

[5] E. Stacey. *Collaborative Learning in an Online Environment*. International Journal of E-Learning & Distance Education, vol. 14, no. 2, pp. 14–33.

[6] H.W. Lin, Wen-Chih Chang, George Yee, Timothy K. Shih, Chun-Chia Wang, and Hsuan-Che Yang. *Applying Petri Nets to model Scorm Learning Sequence Specification in Collabrative Learning*. In: Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05), 2005, pp.203-208. doi: 10.1109/AINA.2005.120.

[7] *Collaborative learning*. [on-line] https://www.valamis.com/hub/collaborative-learning.

[8] *HiPS: Hierarchical Petri net Simulator*. [on-line] http://hips-tools.sourceforge.net/wiki/index.php/Main_Page.

Inga Titchiev[1,2]

[1]Vladimir Andrunahievici Institute of Mathematics and Computer Science

[2]Tiraspol State University
E–mail: inga.titchiev@math.md

Developing augmented artifacts based on learning style approach

Olesea Caftanatov, Inga Titchiev, Veronica Iamandi,
Dan Talambuta, Daniela Caganovschi

Abstract

In this article, the possibilities offered by augmented reality in education are used based on learning style approach in order to increase the efficiency and quality of the study process.

Keywords: Augmented reality, artifacts, marker, learning style, Bartle's player taxonomy, Bloom verb taxonomy.

1 Introduction

Augmented Reality (AR) are technologies which enhance our perception and help us to see, hear, and feel our environments in new and enriched ways. The idea of augmented reality [3] is not new, these technologies have been developed and researched during the last years. Due to different possibilities that it offers, it has become a new trend including the education, but these opportunities are still unsettled and little applied. Thus, in this article, we will research the approach based on the development of augmented artifacts taking into account the learning preferences (learning styles) of the subject, types of students based on Bartle's characters theory. Moreover, for the development of the augmented artifacts, marker-based Augmented Reality approach was used [5], which allows the inclusion of different scenarios, being applicable for different types of learning styles. More about the principle used

in developing augmented artifacts is described in the methodology section. Additionally, while we get excited about various AR features that can be used in the Education field, the image trigger is fundamental to the quality experience. Therefore, we present a few recommendations in designing artifacts and markers in Section 3.1.

2 Background research

2.1 Augmented reality

AR technology has reshaped the way we interact with the real world. Augmented reality is the technology that expands our physical world, adding layers of digital information on it. AR is often mistaken for virtual reality (VR). While they do share pieces of development history, the two are not the same. Unlike VR, augmented reality does not create completely artificial environments to replace the real world with virtual one, it blends technology with the real world. It appears in direct view of an existing environment and adds multimedia elements to it, such as: video, sounds, graphics, etc. AR leaves a little to science and a lot to the imagination. Augmented reality is not a specific device or program, it is a type of human-computer interaction. Augmented reality tech was invented in 1968, with Ivan Sutherland's development of the first head-mounted display system. However, researcher Thomas Caudell coined the term "augmented reality" in 1990 [1].

In the specialized literature, there are various classifications of types of AR. For instance, in [4], the author highlights four types of AR (marker-based, markerless, projection-based and superimposition-based AR), the Wilson's team [15] members think about AR in terms of five distinct experiences types (video launch, 3D object, 360-degree surround, interactive game, information overlay). We identify AR in two major categories: marker-based AR [5] and markerless AR [6], because the other types can be a variation of these two. For our research, we used a marked-based augmentation reality approach to develop artifacts and implement various scenarios in them.

3 Methodology

The research approach develops an augmentation artifact based on five principles:

1. **Marker-based AR**. There are various types of augmented reality; but for our research, we implemented the most common type, which uses markers to trigger an augmentation experience. Due to its use of image recognition, this type of AR is sometimes also called a recognition-based augmented reality.

2. **Bartle's player taxonomy**. There is no field more experienced in engaging users in activity than players in the gaming industry. Every type of player, in our case – students, is unique and special, with their own motivations for engaging. We believe it's nearly impossible to assess and cater to each type of personality. Hence, understanding our audience is important and we need a taxonomy and some kind of assessment system. We decided to classify students based on character theory and player behavior presented in Bartle's taxonomy [18].

3. **VAK learning style**. Having our own preferences for learning, we are all different, this leads to distinct behavioral manifestation [7]. Depending on person, task, context, previous experience, education, etc., behavioral manifestations become over time constant, stable, and frequently applicable, turning into preferences (so-called learning styles) in the learning process. If the teaching material is designed to fit a learning style, then the student who characterizes that style understands the new information better and, consequently, has better results. The most frequently operated in educational practice are the typologies proposed by Barbe, Swassing, and Milone [2], who differentiate the following learning styles (VAK styles); each of these learning styles has particular characteristics [8] as well as learning strategies that can be applied on a case-by-case basis:

(a) **Visual Learners** – prefer viewing written information; transcribing it; using tools for study such as: illustrations, maps, tables, graphs, images, diagrams; emphasizing basic ideas.

(b) **Auditory Learners** – prefer reading aloud; explaining new information, expressing ideas verbally; learning with a tutor or in a group, where they can ask questions, provide answers, express how to understand information orally.

(c) **Kinesthetic Learners** – prefer handling the objects to be learned; arranging the tables and diagrams in the correct order; using movements, dancing, pantomime or role-playing; talking and walking while repeating the knowledge, and applying the learned knowledge in practice.

4. **Bloom verb taxonomy**. Given the fact that, after each lesson taught in schools, high schools, universities, students must obtain some skills, the ideology of the proposed personal learning scenarios concept can be represented by the formula below: S_g–$S_c = S_n$ (1), where S_g represents the general skills, S_c – the current skills, and S_n – the necessary skills to be obtained. It is important to note that the list of competencies is grouped according to Bloom's taxonomy [17].

5. **Geometry 5 grade curriculum**. The material used for augmentation experiences was prepared in accordance with the curriculum for 5 grade, for the geometry subject[16].

Below we present the interaction between artifacts with the mentioned principles (see Fig. 1).

3.1 Designing augmented artifacts and markers

An augmented artifact and marker has quite a few tasks to accomplish. Besides the fact that it has to capture student's attention, entice them to pick up their mobile and scan the image, it should have a high quality

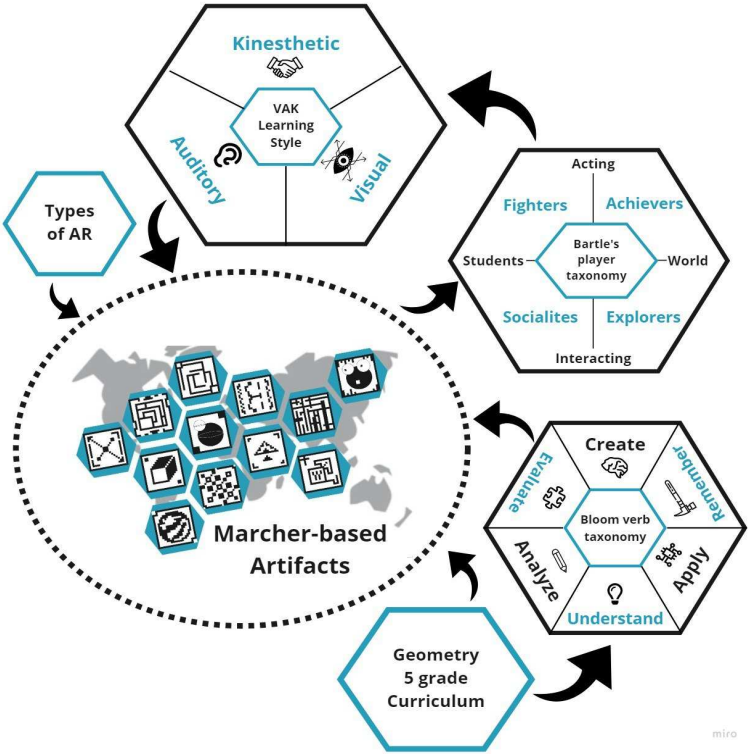


Figure 1. Interaction between artifacts and principles

to let the AR experience come to life. Therefore, in this section, we will describe the best practice to designing augmented artifacts and markers that we observed as a result of the working process and testing. Moreover, we will describe our experiences with low and high star rating image targets.

We consider that **"markers"** are the digital form of image targets that Vuforia Engine can detect and track by comparing extracted natural features from the camera image against a known image target resources database. Markers come in various forms: simple, flat image targets, curled targets in the form of cylindrical shapes, or multi-targets in the composition of a box.

We define **"artifacts"** as the physical form of markers. They can also come in various forms: cards, papers, newspapers, posters, objects, etc. In our cases, it is a laminated image with size 10 x 10 cm. The main purpose of the artifact is to trigger the augmentation content when it is scanned by camera.

When it comes to **designing artifacts**, there are few recommendations on obtaining the best performance from physical target images. Artifacts should be rigid, not flexible. A hard material such as card stock or plastic is better than a simple printed piece of paper, because the flexibility of the printed piece of paper can make it difficult for the object to stay in focus. However, paper artifacts are easily reproducible and widely available, so make sure to fix them to a non-flexible surface.

The size of artifacts varies based on the actual target rating and the augmentation experience. At least 10 or 12 cm (4 – 5 inches) are good for user manipulation. It also depends on the distance between the camera and the artifact; the larger the distance, the bigger your target should be. As an estimate, an artifact with size 20-30 cm wide should be detectable up to about 2-3 meter distance, which is about 10 times the target size. Another important factor is the flatness of the artifact; those that easily bend, coil up, and wrinkle degrade significantly the quality of the tracking. Last but not the least, for attractiveness purposes, the artifacts should be matte but not glossy. Printouts from modern laser printers can be very glossy. Although under ambient

lighting conditions, the glossy surface is not a problem, even so, under certain angles, light sources such as sun, lamp light, etc., can create a glossy reflection. Sometimes the reflection can cover up potentially large parts of the artifacts; as a result, tracking and detection issues can appear. One more recommendation is that when creating artifacts, we should let out our imagination, to make them more creative, relevant, and fun.

When it comes to **designing markers**, there is a range of factors that define how well it tracks when uploaded to Vuforia Target Manager. According to Vuforia Library guideline [14], markers that possess the attributes (see Table 1), will enable the best detection and tracking performance from the Vuforia Engine.

Table 1. Attributes of an ideal image target

Attribute	**Example**
Rich in detail	Street scene, group of people, collages and mixtures of items, and sport scenes are good examples.
Good contrast	Images with bright and dark regions and well-lit areas work well.
No repetitive patterns	Employ unique features and distinct graphics covering as much of the target as possible to avoid symmetry, repeated patterns, and feature-less areas.
Format	Must be 8- or 24-bit PNG and JPG formats; less than 2 MB in size; JPGs must be RGB or greyscale (no CMYK).

There are a few more recommendations to take in account when designing markers. The most important rule here is choosing a unique photo from your collection; however, we decided to create a design from scratch (see Fig. 2). The risk of using stock photos is that someone else may also pick the same picture; as a result, the augmentation experi-

ence may have some issues. For example, when markers are scanned by AR camera, both sides of the AR experience are distorted because the software will recognize the image but will not differentiate the content that has to be retrieved.

Figure 2. Examples of markers version 1.0

Vuforia Engine uses the grayscale version of markers to identify features that can be used for recognition and tracking. If the image has low overall contrast and the histogram of the image is narrow and spiky, it is not likely to be a good target image. Our first batch of markers was designed black-white colors, so after evaluation they were rated 2-3 stars.

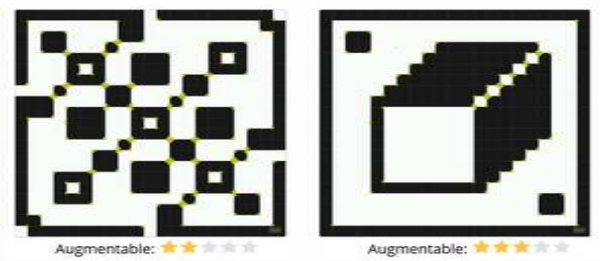


Figure 3. Markers with augmentable rating 2 and 3 stars

An augmentable rating defines how well an image can be detected and tracked using the Vuforia Engine. This rating is displayed in the Target Manager and is returned for each uploaded target via the web API, when using Cloud Reco Databases (see Fig. 3). The augmentable rating can range from 0 to 5 for any given image, where zero indicates that a target is not tracked at all by the AR system.

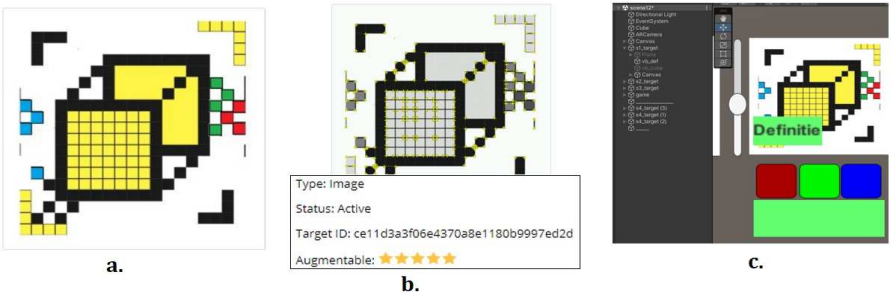


Figure 4. a) marker for square; b) marker evaluations 5 stars through Vuforia Target Manager; c) adding virtual button to marker through Unity platform.

On the other side, the rating of 5, indicates that the marker contains strong detection and tracking ability and is easily recognized by the AR system. Markers with 2-3 rate stars can still be used to trigger simple augmentation content such as visualization of a 3D object, video lessons, text, audio file; but the interaction is almost impossible. For the interaction content, 4-5 star rating is needed. For example, when adding virtual buttons to be easily detected and tracked, it needs to be placed in a zone with many features (see Fig. 4). A feature is a sharp, spiked, chiseled detail in the image, such as the ones present in textured objects. The image analyzer represents features as small yellow crosses; for example, square figure has 4 yellow crosses, circle figure has zero features. Other general features such as organic shapes, round details, blurred, or highly compressed images often do not provide enough richness and detail to be detected and tracked properly.

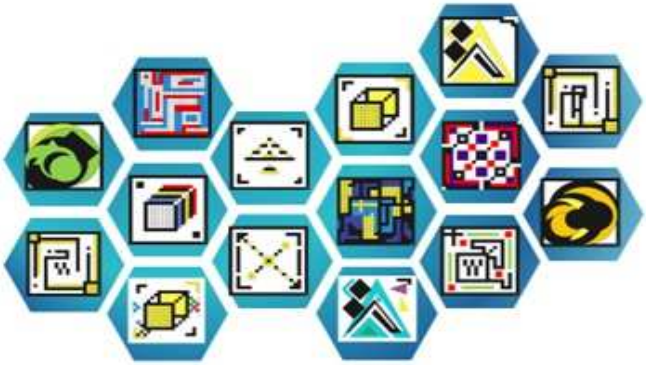


Figure 5. Example of markers version 2.0

After testing the first version of markers, we level up them by adding more color and adjusting contrast; the next batch was created (see Fig. 5). The second version after evaluation got 4-5 rating stars. All features extracted from these images are stored in a cloud database, of which the latter can be downloaded and packaged together with the application (see Fig. 6). The database can then be used by Vuforia Engine for runtime comparisons.

While working with markers, version 2.0, as we observed in some cases, gets a good recognition but a bad tracking during AR experiences. After a few tests, we came to the conclusion that repetitive patterns will confuse the computer vision and will be perceived as the same; it is even more consuming to detect in which direction the marker is placed and to retrieve it. Thus, choosing irregular shapes and photos that look different from all angles, so the computer vision knows if it is upside down per se.

3.2 Results

There are various platforms for creating Augmented Reality experiences, such as Wikitude Studio [9], Bear Go [10], PlugXR [11], etc.; but for our research, we used Vuforia Engine Developer Portal [12] for working with image-based triggers and Unity ver. 2021.3.1.1f1 [13]

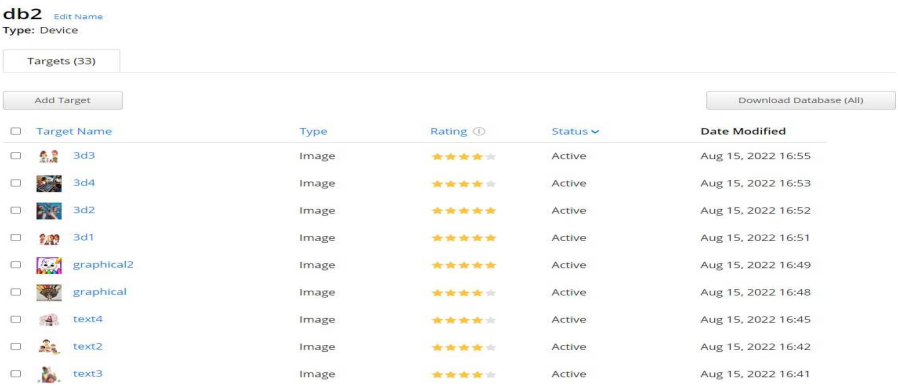


Figure 6. Markers database uploaded and evaluated by Vuforia Target Manager

platform for programming scenarios. In the previous section, we described our work with Vuforia Engine. In this section we will present a working process with Unity platform and a few results.

Using marker-based augmented artifacts, users can interact with the 3D information, objects, and events in a natural way. For example, in Fig. 7, there are presented features to change 3D object size, to rotate, and to change RGB color, or even changing color randomly by pressing a bigger green button.

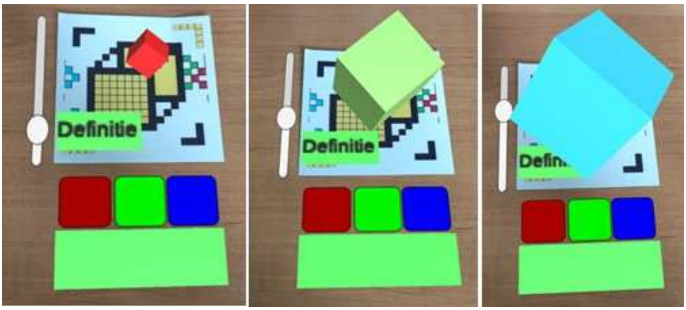


Figure 7. Example of interacting to 3D object

We created 30 types of AR artifacts. When scanned by mobile devices, markers trigger one of augmented experiences, such as 3D objects, video content, audio content, text, formulae, etc. From Bartle's classification, we realized 2 types of experiences for socializers and explorers. For killers and achievers they are still in process. Regarding Bloom's verb taxonomy, we used 30 words, one verb for each artifact (see Fig. 8) for artifacts classification.

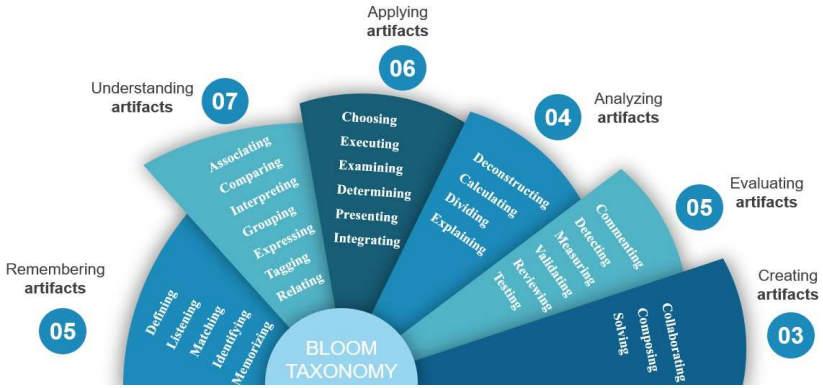


Figure 8. Classifying artifacts based on Bloom's verb used

When developing the scenarios (see Fig. 9) for the artifacts, learning styles based on sensory encoding methods was applied. This will deliver a positive impact by keeping pupils' high engagement and by enhancing their learning abilities like problem-solving, collaboration, imaginative thinking and spatial imagination.

Depending on the predominant sensory organ in receiving information and transmitting it to the brain, the performance obtained in the learning process will be higher.

4 Conclusion

In this article, the process of applying the marker-based Augmented Reality approach was described for the development of augmented artifacts based on learning preferences (learning styles) of the user, types

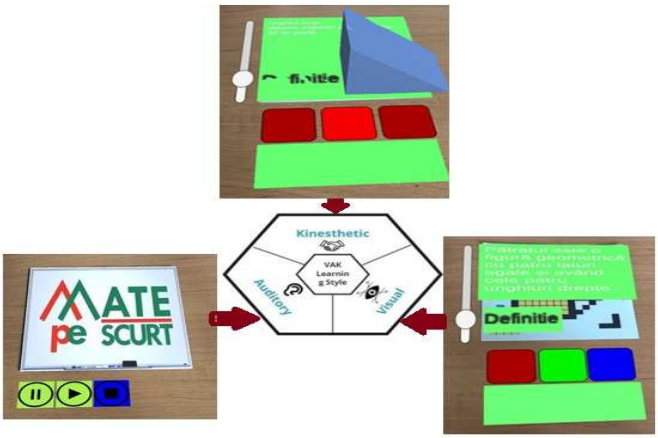


Figure 9. Developed scenarios based on learning style for Augmented artifacts

of students based on Bartle's characters theory, verbs from Blooms taxonomy for artifact classifications. For this, we used Vuforia Engine Developer Portal for working with image-based triggers and Unity platform for programming scenarios. In order to diversify the scenarios in the future work, the other approaches will be applied, such as the markerless one.

Acknowledgments. "Intelligent Information systems for solving ill structured problems, knowledge and Big Data processing" project Ref. Nr. 20.80009.5007.22, has supported part of the research for this paper.

References

[1] T. P. Caudell and D. Mizell. *Augmented reality: An application of heads-up display technology to manual manufacturing processes.* IEEE Xplore Conference: System Sciences, 1992. Proceedings of the Twenty-Fifth Hawaii International Conference on., vol. 2, DOI:10.1109/HICSS.1992.183317, pp 659–669, 1992.

[2] W.B. Barbe, Swassing, and M.N. Milone. *Teaching through Modality Strenghts: Concepts and Practices*, Zaner-Bloser, Columbus, OH, 1979.

[3] D. Schafer and D. Kaufman. *Augmenting Reality with Intelligent Interfaces*, Artificial Intelligence – Emerging Trends and Applications, Chapter 11, pp. 221–242, 2018. DOI:10.5772/intechopen.75751.

[4] Yassir El Filali and Krit Salah-ddine. *Augmented reality types and popular use cases*, International Journal of Engineering, Science and Mathematics, vol. 8, no. 4, pp. 91–97, 2019.

[5] Osman Güler and Ibrahim Yucedag. *Developing an CNC lathe augmented reality application for industrial maintanance training*, Conference: 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–6, 2018. DOI:10.1109/ISMSIT.2018.8567255.

[6] Veronica Teichrieb, Joao Paulo Silva do Monte Lima, Eduardo Lourenco Apolinario,Thiago Souto Maior Cordeiro de Farias Marcio Augusto Silva Bueno, Judith Kelner, and Ismael H. F. Santos. *A Survey of Online Monocular MarkerlessAugmented Reality*, International Journal of modeling and simulation for the petroleum industry, vol. 1, no. 1, pp.1–7, august, 2007.

[7] D. Kolb. *The Kolb Learning Style Inventory*. Version 3. Boston: Hay Group, 1999.

[8] S. Focsa-Semionov. *Invatarea autoreglata. Teorie. Strategii de învățare*, Editura Epigraf, pp.95–116, 2010.

[9] *Wikitude Studio for target images creation*. [Online]. Available: https://www.wikitude.com/products/studio/

[10] *ARGO Developer Portal*. [Online]. Available: https://developer.bear2b.com/

[11] *A Cloud-based Augmented Reality platform*. [Online]. Available: https://www.plugxr.com/

[12] *Vuforia engine, developer portal*. [Online]. Available: https://developer.vuforia.com/

[13] *Unity platform.* [Online]. Available: https://unity.com/

[14] *Vuforia Developer library.* [Online]. Available: https://library.vuforia.com/objects/best-practices-designing-and-developing-image-based-targets

[15] *Trekk a tech-driven creative services agency.* [Online]. Available: https://www.trekk.com/insights/five-types-augmented-reality-experiences-and-how-use-them-your-marketing

[16] *Grade 5 Curriculum.* [Online]. Available: https://www.mathsisfun.com/links/curriculum-year-5.html

[17] *Bloom's Taxonomy.* [Online]. Available: https://www.teachthought.com/learning/what-is-blooms-taxonomy/

[18] Richard Bartle. H*earts, Clubs, Diamonds, Spades: Players who suit muds.* [Online]. Available: https://mud.co.uk/richard/hcds.htm

Inga Titchiev[1,3], Olesea Caftanatov[1], Veronica Iamandi[1], Dan Talambuta[1], Daniela Caganovschi[2]

[1] Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: olesea.caftanatov@math.md
E–mail: inga.titchiev@math.md
E–mail: veronica.gisca@gmail.com
E–mail: dantalambuta@gmail.com

[2] State University of Moldova
E–mail: cda3721@gmail.com

[3] Tiraspol State University

Data in the technologies of modern society

Constantin Gaindric, Galina Magariu, Tatiana Verlan

1 Introduction

The world is going through the fourth industrial revolution, which is an amalgamation of physical, digital, and biological industrial technologies and is creating new opportunities and influences. Success depends primarily on a trained staff, on a coherent system of research and innovation, which will ensure technological and technical skills. No less important is an infrastructure that would ensure the collection, storage, processing, and security of information.

The problems to be solved are systemic and interdisciplinary, which means that the applied solutions must, in turn, be systemic and interdisciplinary. And the first issue, on which success will depend, is the citizen for reforms or reforms for the citizen.

At the Davos forum, among the most impressive achievements expected by 2025 were considered the following: 90% of the population with regular access to the Internet, the first robot with artificial intelligence (AI), a full member of the board of directors of some corporations, 30% of corporate audits will be carried out by AI, tax collection will be carried out through blockchain technologies.

The main problem lies not in these promising ideas, but in their implementation, which requires an information infrastructure, a system of professional education, and research institutions focused on societal problems. In the economy, the main factors of production were considered capital, labor, and raw material. The focus in the current economy, but especially in the future, is on data, knowledge, innovations, and technologies.

All information and analytical systems that are created to help specialists in various fields to make decisions use data. How correct or adequate the decision issued by a particular system will be, depends not only on the qualifications of the system developers and the quality of the resulting software product but to a large extent on the quality of the data used in the system.

And now, when the focus is put on artificial intelligence, and the expectations from an application of AI in many cases are too optimistic, the quality of the data and the trust in the source and the way of their processing become extremely pressing.

Therefore, the problem of data quality is becoming increasingly important in all areas of human activity, not only at the level of production or service units but also in the field of information technology.

2 Definitions for terms "Data" and "Information"

In the modern world including information technology, data is increasingly becoming a key component of the sequence of actions that leads to decision-making. Data is the starting point of the chain of notions *Data-Information-Knowledge*, without which any activity cannot be conceived, and philosophically it is also the foundation on which the *Data-Information-Knowledge-Wisdom* pyramid is built. The process of obtaining, processing, disseminating, and applying data has been critically accelerated; and their volume has also critically increased.

This article reflects some of the results of literature studies that address the problem of data quality and information quality. This problem has been studied for almost half a century. The authors of this article studied more than 30 articles from scientific and specialized literature on this topic written in different periods. It was interesting to analyze the approach that existed in the 20th century and the 21st and compare the changes if any. It should be noted that there is no consensus on the problem of data quality and information quality to this day. However, one can also observe general solutions both in terms

of time (approach to the problem in the past and today) and opinions of different researchers on this topic.

To begin with, it is necessary to define the terms "data" and "information". Very often these terms are used interchangeably, however, there are differences between them, which are explained in the scientific literature. For example, the Cambridge International Examinations manual [1] gives the following definitions:

Data *– a collection of text, numbers, or symbols in raw or unorganized form. They are with no meaning. Data, therefore, has to be processed or provided with a context, before it can have meaning.*

Information *– the result of processing data, usually by computer. This results in facts, which enables the processed data to be used in context and have meaning. Information is data that has meaning.*

Images and sounds can also be attributed to data, since they can be stored and used using a computer (played back and processed). In [2] *"Data are defined as simple facts, either quantitative or qualitative. Information is defined as organized data."*

Article [3] provides the following comment: *"Most definitions refer to a datum as the most basic descriptive element representing a perception or measurement about some object of interest. By itself, a datum's value typically lacks content, meaning, or intent. Information is more than just a set of data; it is the output of a process that interprets and manipulates data into some prescribed format."*

For practical use, it is useful to identify a classification of data into structured and unstructured data. Structured data is that which can be stored in a table, where each object has the same structure (i.e., a set of attributes). Structured data can be easily stored, searched, reordered, and combined with other structured data.

We encounter unstructured data more often than structured data. Unstructured one describes data, where each object in a set can have its own internal structure that is usually different for each object. Differences in structure between individual elements prevent the analysis of unstructured data in its raw form.

We can often extract structured data from unstructured data using artificial intelligence techniques. But the process of transforming unstructured data is expensive and generates significant expenses.

There are two main types of raw data in terms of how they are obtained: collected data and resulting data. The data collected is obtained by direct measurement or observation designed for this purpose. And the resulting data is a product of a process whose purpose is the explanation of the raw data.

One of the most well-known types of resulting data is metadata, that is, data that describes other data.

To better understand the terms "data" and "information" in the field of computer science, some authors explain[1] that data is input, *"or what you tell the computer to do or save"*. But, information can be regarded as output, *"or how the computer interprets your data and shows you the requested action or directive"*.

When describing the key difference between "data" and "information", the author of [6], among several points, gives the following:

- *"Data never depends on Information while Information is dependent on Data";*
- *"Data can be structured, tabular data, graph, data tree whereas Information is language, ideas, and thoughts based on the given data."*

3 About Data quality, Information quality, and criteria for them

The concepts of "data" and "information" are closely intertwined and, in a sense, interrelated (though only in one direction). *"When data are intelligently organized they convey information, and what information is conveyed depends upon just how the data are organized"* [2].

[1] https://examples.yourdictionary.com/difference-between-data-and-information-explained.html

This relationship affects the decision-making process by analysts since it implies not only the competent organization of data but also their quality. If you have bad data, then you will have bad information, which will lead to making wrong decisions, which, in turn, can lead to significant losses in production, business, finances, misdiagnosis, etc. The issue of data quality (DQ) is, to some extent, subjective: data that one user considers to be of high quality may be of low quality to another user. In addition, data that is considered high-quality today will be low-quality data tomorrow. Also, the opposite is possible: data that is considered low-quality data today, tomorrow we can learn how to process it better, and it will become high-quality data.

3.1 Data quality and criteria

There are many approaches to measuring the quality of data and quality of information and defining objective criteria – dimensions. The most popular approach is to answer two questions:

- How suitable is the data to be used for these purposes?
- How well do the data reflect the real situation they describe?

"To increase the trust in data-driven decisions, it is necessary to measure and to know the quality of the employed data with appropriate tools" [4].

Many studies provide some objective criteria for assessing data quality. However, there is still no consensus in the scientific literature on the issue of significant criteria for data quality. In a simplified sense, the quality of data is the degree of their suitability for use. So, for example, in the review [4], the following comment is given:

"DQ is most often associated with the "fitness for use" principle, which refers to the subjectivity and context-dependency of this topic. Data quality is typically referred to as a multi-dimensional concept, where single aspects are described by DQ dimensions (. . .) Our evaluation framework covers the four most frequently used dimensions, which are (. . .) accuracy, completeness, consistency, and timeliness."

ISO 9000:2015 defines the quality of data by the degree to which they meet requirements: needs or expectations, such as completeness, reliability, accuracy, consistency, availability, and timeliness.

Each of these criteria is defined differently in the literature and in different fields of application.

Completeness *"is very generally described as the 'breadth, depth, and scope of information contained in the data' and covers the condition for data to exist to be complete."* [4] Data completeness tells you whether your datasets contain everything you need for your research or management needs.

Reliability is the ability to trust in data used in the organization. The reliability of the data refers to their completeness and accuracy, as well as how consistently the measurements were made, and whether the same results are obtained with the same measurement under the same conditions several times, or by different people.

Accuracy *"can be described as the closeness between an information system and the part of the real world it is supposed to model".* [4] Accuracy means the usage of data that conforms the reality. Moreover, it is very important that the data values are correct not only in their value but in their form too; they should *"be represented in a consistent and unambiguous form"* [2].

Consistency *"captures the violation of semantic rules defined over data items, where items can be tuples of relational tables or records in a file."* [4] As WIKI tells, *"Data consistency refers to whether the same data kept at different places do or do not match."* Also, T.S. Adams in his article gives a very clear and explicative definition[3]: *"Data consistency is the process of keeping information uniform as it moves across a network and between various applications on a computer. There are typically three*

[2]http://etutorials.org/Misc/data+quality/Part+I+Understanding+Data+Accuracy/Chapter+2+Definition+of+Accurate+Data/2.3+Data+Accuracy+Defined/

[3]https://www.easytechjunkie.com/what-is-data-consistency.htm

types of data consistency: point in time consistency, transaction consistency, and application consistency. Ensuring that a computer network has all three elements of data consistency covered is the best way to ensure that data is not lost or corrupted as it travels throughout the system. In the absence of data consistency, there are no guarantees that any piece of information on the system is uniform across the breadth of the computer network."

Availability of data – the term that usually is used in the framework of some organization, institution, or company and means that all data related to this organization are available to its internal goals or its partners at any time 24/7/365. It is essential for the uninterrupted and stable work of this organization and its management without faults.

Timeliness describes *"'how current the data are for the task at hand' and is closely connected to the notions of currency (update frequency of data) and volatility (how fast data becomes irrelevant)"*. [4] Timeliness of the data means that the whole chain "data collecting-transfer-processing-presentation" is run in real-time.

3.2 Information quality and criteria

Similarly, for the question "What kind of information should be considered qualitative" – in different areas of information use, the answer to this question may be different. However, it is important to understand how information is perceived and used by its consumer. There are two necessary steps here [5]:

1. *Highlighting which attributes are important.*
2. *Determining how these attributes affect the customers in question.*

In the specialized literature, 10 attributes of information quality are most frequently used, which can affect the effectiveness of information systems and can contribute to the development of strategies to improve

the quality of information. We will highlight their definitions as follows (see [5], except for "Relevance" with the definition formulated by the authors of this article):

Relevance – how well it reflects the user's needs. If it does not reflect his needs, it still does not mean that the information is bad; it may be good for another class of users.

Accuracy – Accurate information reflects the underlying reality. Less well understood is that information can be too accurate when its degree of precision exceeds its customer's processing capability. This increases the cost of information systems.

Timeliness – The concept of what is timely is itself constantly changing and being redefined, because of changes in customer perceptions caused by technology and the competitive environment.

Completeness – Incomplete information can lead customers astray. However, complete information for one person may be incomplete for another. Information may also be too complete. The danger in business lies in information systems that generate so much information that customers cannot process it all in a timely fashion.

Coherence – is how well the information hangs together and is consistent with itself.

Format – refers to how the information is presented to the customer.

Accessibility – information that can be obtained when needed. It depends on the customer and even on the specific circumstances for that customer. For information quality to occur, timeliness and accessibility should complement each other.

Compatibility – Information quality lies not only in the quality of the information itself, but also in how it can be combined with other information and delivered to a customer.

Security – Two aspects of information security are protecting information from people (logical security) and protecting information from natural disasters (disaster recovery planning).

Validity – Information has validity when it can be verified as being true and satisfying appropriate standards related to' other dimensions such as accuracy, timeliness, completeness, and security.

This set of information quality attributes can change, even within the same domain of use.

Therefore, at each current moment, the user must answer the following questions [5]:

- *First, are yesterday's perceptions of quality needs still valid?*
- *Second, how do quality needs translate into technology requirements?*
- *Third, do internal information collection, dissemination, and verification procedures measure up to quality requirements?*

4 Conclusion

To confidently use data in decision-making in any field, we need not limit ourselves to the attributes described but also need confidence in: the reliability and identification of the supply chain and other attributes of the origin of the data; the recipient's assessment of the public data provider's business qualities as responsible and relatively independent.

Recently, the penetration of AI into all fields is increasingly discussed, but practically very little attention is drawn to the fact that the conclusions and recommendations proposed are based on the data collections used, the quality of which is not evaluated either in terms of veracity or about the sources from where they are obtained.

This gave us an impetus to present some considerations on the issue of data quality and information produced from data processing.

The quantitative evaluation of the whole set (or a smaller number) of significant indicators for a decision-making process allows us to know the degree of suitability for the proposed purpose and to be sure that it will be the expected one.

Acknowledgments. "Intelligent Information systems for solving ill structured problems, knowledge and Big Data processing project"

Ref. Nr. 20.80009.5007.22, has supported part of the research for this paper.

References

[1] *Topic support guide. Cambridge International AS & A Level, Information Technology 9626, For examination from 2017. Topic 1.1 Data, information and knowledge*, Cambridge International Examinations 2015, Version 1, https://www.cambridgeinternational.org/images/285017-data-information-and-knowledge.pdf.

[2] Sue P. Stafford. *Data, Information, Knowledge, and Wisdom*, in: Knowledge management, organizational intelligence and learning, and complexity, vol. III, Encyclopedia of Life Support Systems (EOLSS), 2011.

[3] Tuomo Uotila and Helina¨ Melkas. *Quality of data, information and knowledge in regional foresight processes*, Futures 39, 2007, pp. 1117–1130.

[4] Lisa Ehrlinger, Elisa Rusz, and Wolfram Wöß. *A survey of data quality measurement and monitoring tools*, Preprint, arXiv:1907.08138v1 [cs.DB], 18 Jul 2019.

[5] Holmes E. Miller. *The multiple dimensions of information quality*, Information Systems Management, vol. 13, no. 2, pp.79–82. DOI: 10.1080/10580539608906992.

[6] David Taylor. *Difference between Information and Data*, Updated July 9, 2022, Available: https://www.guru99.com/difference-information-data.html

Constantin Gaindric[1], Galina Magariu[2], Tatiana Verlan[3]

Vladimir Andrunachievici Institute of Mathematics and Computer Science
5, Academiei street, Chisinau, Republic of Moldova, MD 2028
E–mails: [1]constantin.gaindric@math.md
[2]galina.magariu@math.md
[3]tatiana.verlan@math.md

Artificial Intelligence Strategies: Republic of Moldova relative to European Union countries

Svetlana Cojocaru, Constantin Gaindric, Tatiana Verlan

Abstract

The paper examines the policies of European countries regarding the use and development of applications based on Artificial Intelligence. Along with the examination of the strategic documents of the European Commission, some of them are studied at national level. Also, the situation in this field in the Republic of Moldova is analyzed.

Keywords: Artificial intelligence, strategy, education, research and development.

1 Introduction

Artificial intelligence (AI) becomes the driving force of the digital age. AI-based applications are used more frequently, often without this fact being explicitly realized. Automatic translation, the quality of which is getting better, and also contextual advertising are just two of such examples that each of us knows. Among the objectives of the new Digital Europe 2021-2027 Program, there is the massive implementation of solutions based on artificial intelligence, especially in critical areas such as climate change or health.

In Section 2 of this article, we will examine the evolution of the definition of artificial intelligence, as well as the basics of the strategy adopted by the European Union. Community countries, in turn, have developed their own national strategies (or they are in the process of developing them). In Section 3 we will examine the specifics of the approaches in three countries: Estonia, Bulgaria, and Romania. The topic of Section 4 is the situation in the Republic of Moldova

examined from the point of view of human potential, infrastructure, existing developments that could be the subject of public and private sector implementations, as well as the reflection analysis of the usage aspects of solutions based on artificial intelligence in the government program, adopted in August 2021.

2 Some definitions of the term "Artificial Intelligence"

In our contemporary life, the term "Artificial Intelligence" (AI) is becoming common and more and more frequently used. Meaning of this term evolutionized since the time of its first mentioning. Depending on the area of its application, the definition of AI focuses on different subtleties, and also something is added over time and with the development of modern information technologies. For the first time it was used by John McCarthy, the American computer scientist, in 1956 at a summer seminar at Dartmouth College (Hanover, USA). He defined AI [1] as *"the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence."*

Definitions of the term "Artificial Intelligence" in different kinds of literature explain its essence depending on the goal for which this term is applied, but they are not contradictory.

The compilers of the explanatory dictionary on artificial intelligence tried to collect and systematize special terminology on artificial intelligence and intelligent systems [2]. So, they define AI in two parts:

> *"1. A scientific direction, within the framework of which the problems of hardware or software modeling of those types of human activity that are traditionally considered to be intellectual are set and solved.*
>
> *2. The property of intelligent systems to perform functions (creative), which are traditionally considered the prerogative of a person."*

After Oxford English dictionary, Artificial Intelligence [3] is *"The theory and development of computer systems able to perform tasks nor-*

mally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages."

IBM Cloud, a company that offers the most open and secure public cloud for businesses, a next-generation hybrid multicloud platform, advanced data and AI capabilities, and extensive experience with large enterprises across 20 industries, gives the following definition for AI [4]: *"Artificial intelligence is the use of computers and systems for the simulation of the human mental process to solve problems and make decisions."*

NetApp, Inc., an American company, one of the five world leaders in the market of disk storage systems and solutions for storing and managing information, in its article about AI, gives its definition and explains some scenarios of Artificial Intelligence Use Cases [5]: *"Artificial intelligence is the foundation for simulating human intelligence processes by creating and applying algorithms embedded in dynamic computing environments. Simply put, AI is trying to make computers think and act the way humans do. Achieving this goal requires three key components: Computing systems, Data and data management, Advanced AI algorithms (code). The closer the desired result is to humans, the more data and processing power is required."*

The comprehensive definition is given by the independent High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission [6]:

> *"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors, and actuators, as well as the integration of all other techniques into cyber-physical systems).”

The AI HLEG proposes to use this definition when projecting strategy for AI development.

The White Paper on AI created by the European Commission (EC) emphasizes the fast development of AI in modern society. AI penetrates our daily life and solves various problems: in the field of financial services (e.g., detecting fraud), interacting with customers online (e.g., online chatbots), speech recognition (e.g., in mobile devices), computer vision (e.g., face recognition in photos, analysis of medical X-rays), analysis of data on customer behavior to predict their further purchases, systems for mass consumption goods selection subject to the preferences of different consumer groups, smart home systems and household robot assistants, Internet search systems, maps and location determination, and much more.

Nevertheless, there are also serious risks to which AI usage can give rise. One of these risks is the issue of personal data protection. Another is that the reliability of the results is highly dependent on data that may be incomplete, biased, or of poor quality. The European Commission in its White Paper [7] notes the following risks: “... opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes”. So, the EC maintains regulatory measures and funding possibilities for a dual purpose: on the one hand, to promote the AI implementation; on the other hand, to pay attention to potential risks that this new technology may bring: “This White Paper presents policy options to enable a trustworthy and secure development of AI in Europe, in full respect of the values and rights of EU citizens.”

Another regulatory document of the European Commission, Coordinated Plan on Artificial Intelligence, “puts forward a concrete set of joint actions for the European Commission and Member States on how

to create EU global leadership on trustworthy AI" [8]. First of all, in this document, there are indicated three main conditions to achieve the set goals: 1) appropriate governance and coordination framework; 2) data (large, high-quality, secure, and robust datasets); 3) computation infrastructure (necessary for storing, analyzing, and processing the increasingly large volumes of data).

3 Case study: approaches to the artificial intelligence use in different European countries

In this section, we will examine the examples of implementation of the Coordinated plan on artificial intelligence [8] (2021) in several EU countries. Multiple international reports [9-11] analyze state of the art in the field of national AI strategies in terms of the following policy areas:

- **Human capital.** This section includes policies aimed at educating the population of all ages in the field of using and developing AI-based solutions. They cover both the respective courses in the programs of educational institutions of all levels, as well as refresher trainings for employees of different specialties to cultivate their skills needed to operate with AI-based systems.
- **From the lab to the market** policies are inclined to support research and innovation in AI in order to assure business growth in the private sector and increase the efficiency of public services.
- **Networking** refers to collaborations in the field of AI promoted by the private and public sectors, including those at the international level.
- **Regulation** covers policies related to ethical issues, the regulatory framework, the adoption of international standards.
- **Infrastructure** refers both to the development aspects of digital and telecommunications infrastructure itself and to solving the

problems of data collection, use, and sharing at the national and international level.

In our examples, we will refer mainly to the first two aspects.

The first example is that of Estonia. We chose this country for our case study for two reasons:

- its success in digital transformation is well-known;
- it is one of the former Soviet countries, so we could say that after the declaration of independence in August 1991, the Republic of Estonia and the Republic of Moldova were at the same starting point.

The Government of Estonia adopted its National Artificial Intelligence Strategy in July 2019. The document [12] envisaged the implementation of a series of actions divided into four compartments:

1. Advancing the uptake of AI in the public sector in Estonia;
2. Advancing the uptake of AI in the private sector in Estonia;
3. Developing AI R&D and education in Estonia;
4. Developing a legal environment for the uptake of AI.

Most actions (30 in number) are provided concerning the public sector. Along with those of information, involvement of public authorities, trainings, etc. we will also mention the specific presence in this compartment of the R&D projects related to the implementation of automatic AI-based decision-making support in Estonian state institutions. Within this project [13] several applications based on artificial intelligence approaches have been developed, including:

- A system to support NEET youth, which is made available to municipal workers in the field of social assistance, especially those dealing with child protection and youth problems, to identify and support young people aged 16-26 who are Not in Education, Employment, or Training.

- Prediction model for the healthcare needs of patients with chronic illnesses. The system, implemented in the pilot version and based on machine learning algorithms, is made available to family physicians by assisting in identifying patients on their list with multiple chronic illnesses who would benefit most from additional help with prevention, counseling, and follow-up care to improve their quality of life.

- Machine learning software to match job seekers with employers. Based on the European Skills, Competences, Qualifications, and Occupations (ESCO) classification system, developed by the European Commission, which defines skills needed in many areas of life, there was elaborated the machine learning algorithm which chooses candidates with skill categories suitable for the corresponding job profile. At the time of reporting (2020), the system operated with over 400,000 user profiles (remember that the population of Estonia is 1 million 325 thousand inhabitants) and a smaller number of workplace profiles, and the hiring process could be fully automated and did not take more than 5-10 days.

- Machine vision AI solution for better traffic management. The system is implemented in Tallinn and serves to monitor road traffic, especially public transport. Based on the information collected and processed within the system, it is possible, along with other aspects, to make decisions about parking problems or road construction.

We will also mention the special actions in the field of R&D, within which three relevant research groups are funded: AI and machine learning, data science and big data, robot-human cooperation. For the research and elaborations carried out by these groups, 1.5M EUR are allocated annually.

A series of activities are also envisaged in the field of education, the leader being the University of Tartu. An important role belongs to the IT Academy program (English brand name StudyITin.ee), which includes the collaboration of the state, educational institutions, and ICT companies to ensure an advanced quality of studies and research

in the field of ICT. It involves training about 50 master's students specializing in AI at the University of Tartu in the period 2020-2023, reviewing the curricula of courses in general schools with the inclusion of AI subjects.

Last but not least, we would like to point out that the AI Program in Estonia is called the "Kratt plan". Kratt is a character from local mythology, an artificial creature, who serves his master by performing various works, which the master orders him to do. The need to pay tribute to the devil for the creature to function also alludes to some ethical issues, which are intensely discussed concerning the vast application of AI in various fields.

From the above, we can conclude that this country has all the chances to achieve its ambitious goal formulated as follows: "Estonia could become the role model and testbed for the rest of the world as a place where Kratt, or AI, is put to work for the people's well-being in both the public and private sectors" [14].

Another example, which we will examine, is that of Bulgaria. We have selected this country thanks to several similarities, which we can observe, as we will see from the following, as compared with the situation in the Republic of Moldova. The country's strategic document is entitled "Artificial intelligence for intelligent growth and a prosperous democratic society", the project was developed by the Bulgarian Academy of Sciences and approved in December 2020. The concept envisages the realization of a series of actions over the next ten years, based on existing results in the field of artificial intelligence uptake and the development of AI-based applications [15].

Several sources (such as [16]) mention the existence of over 50 companies working with artificial intelligence applications, the most important areas in this regard being retail, finance, and media. The research and elaborations in the field of natural language processing also have international recognition. In [17] it is emphasized that the main areas of specialization in Bulgaria are big data, predictive analytics, data science, and chatbots. According to the same source, the share of the IT sector in GDP formation in Bulgaria is 3.4%, which is very close to the figures in Moldova: according to data from 2015-2019, the IT industry has reached 3.1% of GDP [18].

The strategy document stipulates that a fundamental proposal for Bulgaria is "focusing on technological specialization in the field of data economy, as the country would have difficulty when realizing strong industrial specialization due to the lack of a critical mass of top industrial companies in the AI sector. Today, the trend is for data to come to the fore in AI and for the emphasis on automatic self-learning to shift from algorithms to data" [15].

The following domains and directions for AI development and implementation are established as the priority ones:

- Software industry;
- Creating AI applications for educational purposes;
- AI applications in public services;
- Intelligent agriculture;
- Applications of AI in healthcare and medicine;
- Applications of AI in ecology and environment.

Special attention in the strategic concept is paid to the education and research domains, namely here we find several features, specific to our country too. Whereas work in the IT sector is much better paid than research, most young people either do not want to start or begin and leave their careers in universities and research institutes, preferring to work in IT companies in the country or abroad.

The basic recommendation for Bulgaria is the need to overcome the fragmentation between small units that develop AI and creation of the conditions for building human potential in a connected national academic environment. Thus, it is proposed:

- Establishment of a Bulgarian center of excellence in AI, which will unite scientific organizations and universities with proven achievements in the field of AI research;
- Involvement of Bulgarian research teams in European artificial intelligence and digitization networks;
- Inclusion of Bulgarian research teams in European testing and experimentation centers related to healthcare, robotics, and agriculture.

In the case of Romania, it is also necessary to mention the initiative of the researchers from the Romanian Academy, who in October 2019 published a document entitled "Manifesto for adaptation to the digital age" [19]. The document specifies the favorable factors for Romania, namely:

- the weight of the ICT sector in GDP (according to the 2019 Country Report, published by the European Commission [20], it was 6-7%);
- Internet infrastructure with high traffic speeds;
- wide penetration of mobile devices among the population;
- the special receptivity of young people towards these technologies.

In the recommendations elaborated by the authors of the "Manifesto", several directions of action are established, in which an important role belongs to research, education, public administration, media.

Although Romania does not yet have a finalized national strategy (in [10] it is mentioned that its elaboration is "in progress"), the Romanian Digitization Authority (RDA) considers that "Artificial Intelligence can revolutionize the activity of public administration", thereby offering "better public services, safer transport systems, personalized products and services that are cheaper and more sustainable" [21].

4 Some aspects of promoting AI in the Republic of Moldova

The importance of information technologies has been realized in the Republic of Moldova since the 90s of the last century. Our country was among those that in 1990 included in the structure of its Government a Ministry of Informatics, Information, and Telecommunications. Subsequently, it underwent several changes in both name and duties, but regardless of them, ICT development and implementation policies were constantly promoted, thanks to which it was possible to create a communication infrastructure based on optic fibers, which had good coverage in the country; there have been implemented services intended for

citizens and economic agents, based on digital technologies; measures have been carried out to equip schools with computers and connect them to the Internet (Program SALT, adopted in 2004, assumed the maintenance of physical access to the Internet for all schools of the country [22]).

Today, according to the analysis of the ICT sector involvement in the economy of the Eastern Partnership countries, carried out by the German Economic Team [23], in the respective sectors of Armenia, Belarus, Georgia, and Ukraine, the ICT revenues in 2019 accounted for 7.1% of GDP. In Moldova, there is observed a share of 5% of GDP, and exports of ICT services represent more than 15% of services exports and about 6.5% of total exports, but only 2% of Moldova's GDP, while ICT infrastructure is very well developed and the number of users with broadband internet access has increased significantly.

The conditions for the IT sector development are good, because there is a developed infrastructure and the population uses ICT technologies extensively, and the companies' expenses are increasing according to the National Bureau of Statistics from 500 thousand lei in 2013 to 2500 thousand lei in 2019.

If in 2005 the index of internet penetration in Moldova was 7.4% compared to 35.5% in Europe, currently according to "The future of IT Landscape Report. The ultimate guide for IT buyers and investors looking to source in emerging Europe", developed by Emerging Europe in 2021 [24], the internet penetration rate in Moldova is already 76% with an increase of more than 10 times compared to 2005.

Thanks to higher salaries, the opportunities offered by ICT are attractive, and yet the share of the workforce in the ICT sector is relatively small.

Based on the data brought up, let's see how the field of AI in the Republic of Moldova is presented in the light of EU regulatory documents. We will mention that by June 2021, 20 EU Member States and Norway had published their national AI strategies, while 7 Member States were in the final drafting phase.

On August 4, 2021, the Parliament of the Republic of Moldova approved the Government Activity Program "Moldova of Good Times" [25], which also contains a section dedicated to digital transformation.

In this section as well as in the others, the notions of artificial intelligence, intelligent municipality, intelligent instruments, etc. are encountered. The Program stipulates that "The state must be able to capitalize on the opportunities offered by the digital revolution, but also to manage the risks generated by it". However, the program does not explicitly state, as in other countries' policy documents, that artificial intelligence will influence the increase of efficiency of services provided to citizens by state authorities, relaunch industry, streamline agriculture, mitigate climate change, and improve healthcare. The provisions of the program are limited to "Studying and exploitation of initiatives and programs of EU countries in the field of adopting artificial intelligence technologies, robotics, blockchain, smart contracts and other emerging technologies to modernize public and private digital infrastructures with the purpose to deliver better services, operational effectiveness, and strengthening of the country's cybernetic capacity".

Thus, the government intends to use modern working tools by intensifying the application of information technologies to exclude the flow of paper documents in administrative processes. On the other hand, the White Paper [7] rightly states: "Europe's current and future sustainable economic growth and societal wellbeing increasingly draws on value created by data. AI is one of the most important applications of the data economy". In this context, the digital inclusion of local authorities is envisaged by creating a digital platform with access to centralized information resources, and public services to be rethought and modernized with a focus on the citizen.

The continuation of the Government Services' modernization, taking into account the Government's vision expressed in the Program, based on the Action Plan concerning public service modernization reform, would capitalize on and continue the achievements reached in the framework of the ongoing e-Government Transformation Project. Also, it would contribute to a) reorganizing public administrative services for the purpose to be provided implicitly and electronically as a priority, with the result of the service delivered in the form of an electronic document; b) increasing access, efficiency, and quality in the provision of government services. A key element of success is the evaluation of the quality and accessibility of services by beneficiaries – citizens, because

a considerable part of them does not trust the quality and safety of electronic services. And another aspect is the use of the set of artificial intelligence technologies that combine data, algorithms, and ascending dynamics of internet penetration.

To have an overall view of national policy initiatives and national AI strategies, we will return to their examination in the light of the five policy domains, presented in the previous section:

- **Human capital.** There are currently five universities in the Republic of Moldova, where ICT specialists are being educated, but there is no master's program specifically aimed at AI.
- Ways and means of **passing from the laboratory to the market**. As it was mentioned earlier, this compartment includes policy initiatives to encourage research and innovation in AI for business growth in the private sector and public services' efficiency increase. Tools are also included to facilitate the experimentation of AI pilot products and newly developed services.

The propulsion of new AI products from the laboratory to the market can only succeed in an enterprise-based environment, with funding for research and innovation in AI, which would support the transformation of AI concepts into successful products and services. Mechanisms are needed for the adoption and use of AI in public administration. In this context, EU countries have taken measures to stimulate AI research and have developed or are in the process of setting up national centers of competence in AI research. Some centers are aimed at many domains of research, others are focused on autonomous systems, cyber security procedures for AI systems, machine learning, Data Science. The most frequently reported sectors in national strategies are agriculture, healthcare, transport, and power engineering.

We will make a brief overview of the Republic of Moldova's research related to AI. From the analysis of different countries' national strategies, we can see that several of them have given priority to AI language technologies for interactive dialogue systems and virtual assistants for personalizing public services. Denmark, Norway, Portugal, Slovakia, and Spain have included support policies for research in natural language processing. The research carried out in the Republic of Moldova

in natural language processing is in line with EU visions, and the results obtained over the years are at the European level. In particular, we will emphasize the achievements in the recovery of the country's cultural heritage. The systems for digitizing old texts (starting with the 17th century), developed within the "Vladimir Andrunachievici Institute of Mathematics and Computer Science", allow the restoration of works of historical value in the wide circuit, offering specialists in various fields and the general public as well a tool for accessing these printings in a convenient, editable format, in an original or contemporary script.

The project "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" is in the trend of EU-recommended research. Several national projects and those of the partnership with EU countries have aimed to develop medical information systems oriented to providing physicians with support in the diagnostic process, as well as modeling the course of patient treatment. The formation of a Network for informational methods in supporting persons predisposed to preventable strokes using common devices has been initiated. The international project "An Adaptive Decision Support Framework for the Management of Mass Casualty via an Artificial Intelligence Based Multilayered Approach integrating an Intelligent Reachback Information System" is underway, in which, together with the Republic of Moldova team, researchers from Germany, Croatia, Romania, USA participate. Researchers from our country have joined some COST Action projects, such as "European Network for Combining Language Learning with Crowdsourcing Technique", "A Network for Gravitational Waves, Geophysics and Machine Learning", etc.

The problem of AI is of importance that should not be neglected, a fact that requires the creation of a national center of competence in AI research, because the number of researchers in research institutes and universities is far below the limit of necessities, being worthy of following the experience of several EU (but not only) countries, which directed AI research in national programs to the needs of countries, without forsaking the general perspective directions.

The innovation and usage of AI in public administration are stimulated, including AI programs for public services, e-government strategies to improve the digitization of public administration processes, pub-

lic procurement, and exchange of good practices. The e-government center at the Government of the Republic of Moldova managed to offer the society a series of services, the last being MDelivery, which significantly changed the processes and time to obtain some documents for which citizens previously lost time and effort, but also, which is more important, have to some extent influenced the attitude of the population towards the government.

For the time being, a greater interest in the application of AI-based technologies in our country is attested in the finance & banking and insurance sectors. The National Bank of Moldova is among the institutions that, operating with such systems, will have the possibility of interconnected supervision of all banking operations, as it will function as a single point of access and visualization of information contained in numerous databases. The concept of operation is based on risk analysis, which will make it possible to detect suspicious activities and issue alerts, early identification of risks of money laundering and terrorist financing, and the detection of suspicious changes in the ownership structure of banks [26].

- **Networking** comprises the set of virtual means and AI collaboration initiatives in the private and public sector, including those with foreign people and companies. Networking includes dissemination policies, promotional campaigns, and mapping of AI applications. Many governments have put in place policies to build innovation communities by bringing together technology companies, research centers, and innovation actors. Many countries also set policies to attract skills and investment from AI abroad. In this respect, some countries have dedicated strategies, such as the researcher mobility program in Cyprus and the future Spanish Talent Hub program. Other policies aim to improve working conditions for foreign talent by facilitating administrative procedures. The Czech Republic, Finland, Italy, Malta, Portugal, and Spain are implementing this through starting visas and fast services for valuable talent coming from abroad.

  Our country has other problems, namely the loss of young specialists who either emigrate to countries that provide them with a

well-paid and interesting job or work in foreign companies, which have other objectives than their home country.

Most countries exploit social channels to raise awareness in respect of AI and increase networking opportunities. Slovenia intends to launch a communication platform for the collection and dissemination of good practices and case studies on the use and implementation of AI in society. Hungary announces the annual award for innovations and AI application projects. Therefore, the Republic of Moldova can also take over a series of good practices in this area.

- **Regulations** provide policies that address issues that refer to human rights, confidentiality, fairness, algorithmic prejudice, transparency and explicability, security and responsibility, etc. To facilitate the development of ethical guidelines, many governments have formed AI ethics committees or councils. These bodies are tasked with developing recommendations on ethical problems and monitoring the use and development of AI technologies. Slovakia is preparing a new act in respect of data to better define data protection regulations, data access principles, and open data regulations. Finland and Portugal are developing national regulations for determining liability problems. The Moldovan Parliament has already adopted a package of laws: 1) Digitization of the economy, Package I, which will facilitate remote interaction in digital format between the Government, business and consumers; 2) Package II is in the process of being worked on; 3) A draft law on public services is also in the process of being examined, which implies that public services will be provided in electronic form as a priority. However, according to [27], Moldova lacks specific regulations on new digital technologies, such as Artificial Intelligence or blockchain.
- **The infrastructure** focuses on the problems of digital and telecommunications infrastructure development and provides initiatives to encourage data collection, use, and sharing. Since AI algorithms imply large amounts of data, it is crucial to establish an environment conducive to infrastructure development to ensure reliable, high-quality data that can be shared with users

in an accessible and robust way. As it was mentioned above, Moldova has a relatively good infrastructure, providing fast connectivity with country-wide coverage. However, there are still discrepancies between urban and rural areas and between population groups with different income levels.

Several EU Member States have drawn up national strategies to lay the foundations for the use and exchange of data that describe the actions needed for open data governance, the creation of data warehouses, the improvement of data interoperability, and the protection of individual and collective rights.

Open data platforms and portals have been developed in all EU Member States, Norway, and Switzerland. They usually aim to provide free access to public administration data.

In this direction, vigorous measures are required from the e-government service of the Government of the Republic of Moldova.

These policy areas are in line with the actions proposed in the Coordinated Plan on Artificial Intelligence [8] and with the policy recommendations addressed to governments, contained in the OECD Recommendation on AI [28].

5 Conclusion

The European Union, as the main goal, has proposed massive implementation of digital technologies in enterprises, putting them at the service of citizens and public administrations. Analyzing the policy documents related to AI in different countries, we note that an important role is assigned to education and research & development. Less visible this aspect appears in the program documents of the Republic of Moldova. Along with the actions envisaged for study and exploitation of EU countries' initiatives and programs on artificial intelligence, as well as the interaction with the Ad hoc Committee on Artificial Intelligence of the Council of Europe, the study on national strategies in other countries indicates the need for active involvement of researchers, but also their support from the state for the achievement of a comprehensive and efficient digital transformation.

This paper is the extended and revised version of the conference paper [29] presented at WIIS 2021.

Acknowledgments. The research was supported by the project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge, and big data".

References

[1] J. McCarthy, "What is Artificial Intelligence?," *Computer Science Department Stanford University*, Stanford, 2004 Nov 24. https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf.

[2] A. N. Averkin, M. G. Gaaze-Rapoport, and D. A. Pospelov, *Explanatory Dictionary of Artificial Intelligence*, Moscow: Radio and communication, 1992, 256 p. http://www.raai.org/library/tolk/aivoc.html#L208.

[3] "Definition of artificial intelligence," *Oxford University Press*, Lexico.com, 14 July 2021. https://www.lexico.com/definition/artificial_intelligence.

[4] IBM Cloud Education, "Artificial Intelligence," June 3, 2020, https://www.ibm.com/ru-ru/cloud/learn/what-is-artificial-intelligence.

[5] NetApp, "What is Artificial Intelligence?," https://www.netapp.com/ru/artificial-intelligence/what-is-artificial-intelligence/.

[6] Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, "A Definition of AI: Main Capabilities and Disciplines," 8 April 2019.

[7] European Commission, "White Paper On Artificial Intelligence – A European approach to excellence and trust," Brussels, 19.2.2020.

[8] European Commission, "Coordinated Plan on Artificial Intelligence 2021 Review," Brussels, 21.4.2021.

[9] V. Van Roy, "AI Watch – National strategies on Artificial Intelligence: A European perspective in 2019," EUR 30102 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-16409-8, doi: 10.2760/602843, JRC119974.

[10] V. Van Roy, F. Rossetti, K. Perset, and L. Galindo-Romero, "AI Watch – National strategies on Artificial Intelligence: A European perspective," 2021 edition, EUR 30745 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-39081-7, doi: 10.2760/069178, JRC122684.

[11] "State of implementation of the OECD AI Principles: Insights from national AI policies," OECD Digital Economy Papers, No. 311, OECD Publishing, 2021, Paris, https://doi.org/10.1787/1cd40c44-en.

[12] "Estonia's national artificial intelligence strategy 2019-2021," https://f98cc689-5814-47ec-86b3-db505a7c3978.filesusr.com/ugd/7df26f_27a618cb80a648c38be427194affa2f3.pdf.

[13] Fabio Chiusi, Sarah Fischer, Nicolas Kayser-Bril, and Matthias Spielkamp (eds.), "Automating Society Report 2020," AlgorithmWatch gGmbH, Berlin, Germany, 298 p. https://automatingsociety.algorithmwatch.org/.

[14] "National AI strategy for 2019-2021 gets a kick-off," https://e-estonia.com/nationa-ai-strategy/.

[15] "Development concept of artificial intelligence in Bulgaria until 2030: Artificial intelligence for smart growth and a prosperous democratic society," https://www.mtitc.government.bg/sites/default/files/koncepciyazarazvitienaiivbulgariyado2030.pdf. (in Bulgarian).

[16] *Artificial intelligence ecosystem in Bulgaria*. https://investsofia.com/wp-content/uploads/2019/06/Artificial_intelligence_ecosyste m_in_Bulgaria_2019-SeeNews_and_Vangavis.pdf.

[17] G. Angelova et al., "Role of education and research for artificial intelligence development in Bulgaria until 2030," in *Mathematics and Education in Mathematics. Proceedings of the Fiftieth Spring Conference of the Union of Bulgarian Mathematicians*, 2021, pp.71–82.

[18] "The ICT sector has become one of the locomotives of economic growth in the Republic of Moldova," https://www.moldpres.md/news/2020/09/03/20007029. (in Romanian).

[19] "Manifesto for adaptation to the digital age," https://acad.ro/media AR/com2019/c1016-ManifestEraDigitala.htm. (in Romanian).

[20] "Country Report Romania 2019 Including an In-Depth Review on the prevention and correction of macroeconomic imbalances," https://ec.europa.eu/info/sites/default/files/file_import/2019-european-semester-country-report-romania_en.pdf.

[21] "ADR launches the public consultation process on the use of artificial intelligence in Romania," https://www.adr.gov.ro/adr-lanseaza-procesul-de-consultare-publica-privind-utilizarea-inteligentei-artificiale-in-romania/. (in Romanian).

[22] L. Burtseva et al., "Digital divide: A glance at the problem in Moldova," in *Information Communication Technologies: Concepts, Methodologies, Tools, and Applications*, IGI Global, Vol. IV, 2008, pp. 2531–2565.

[23] Carolin Busch, Ricardo Giucci, "Moldova's ICT sector in comparison to Ukraine, Belarus, Georgia and Armenia, 2021," https://www.german-economic-team.com/moldau/wp-content/uploads/sites/4/GET_MDA_PB_04_2021_EN.pdf.

[24] https://emerging-europe.com/it-landscape-report/.

[25] https://unpaspentru.md/2021/08/03/program-de-activitate-al-guvernului-moldova-vremurilor-bune/.

[26] https://www.bnm.md/ro/content/premiera-regiune-moldova-va-implementa-cu-suportul-usaid-o-solutie-it-de-ultima-generatie.

[27] "Digital Readiness Analysis. Moldova," UNDP Report, 2021. https://www.md.undp.org/content/moldova/en/home/library/effective_governance/digitalreadinessassessment.html.

[28] OECD, "Recommendation of the Council on Artificial Intelligence," OECD/LEGAL/0449, 2019. https://www.fsmb.org/siteassets/artificial-intelligence/pdfs/oecd-recommendation-on-ai-en.pdf.

[29] Svetlana Cojocaru, Constantin Gaindric, Tatiana Verlan, "Considerations on the Artificial Intelligence Strategies," in *Workshop on Intelligent Information Systems (WIIS2021)*, (Chisinau, Republic of Moldova), October 14-15, 2021, pp. 89–111.

Svetlana Cojocaru[1], Constantin Gaindric[2], Tatiana Verlan[3]

Vladimir Andrunachievici Institute of Mathematics and Computer Science,
5, Academiei street, Chisinau, Republic of Moldova, MD 2028

[1] E–mail: svetlana.cojocaru@math.md

[2] E–mail: constantin.gaindric@math.md

[3] E–mail: tatiana.verlan@math.md

Part 2

Platform for the digitization of heterogeneous documents

On convergent technology in development of information systems for processing of documents with heterogeneous content

Alexandru Colesnicov, Svetlana Cojocaru, Ludmila Malahov, Lyudmila Burtseva

Abstract

Recognition of heterogeneous documents is a challenging problem. To solve it, a toolkit is proposed in the form of a platform for processing documents containing heterogeneous content. For the development, we used convergent technology, according to which both existing software tools, as well as those developed by us, can provide all stages of the recognition process. The development of a platform for solving the problem is done in a virtual environment. The paper describes a functional subset of the platform for solving one subtask: to analyze a scan of a document and to cut it into segments.

Keywords: Recognition of heterogeneous documents, convergent technology, virtualization.

1. Introduction

Development of information technologies led to a variety of specialized systems and software packages for solving specific problems. Currently, there are problems that can be solved using convergent technologies, which represent convergence of existing material, information and cognitive technologies.

We define convergent technologies through the general principle of physical and/or logical pooling of resources based on any resemblance, belonging on similarity of tasks being solved in some aggregate or document processing platform, resource pools, communications, systems and/or data storage or backup networks.

During document digitization, we met the necessity to use for processing the tools that provide support for recognition not only of the text but of other elements of heterogeneous nature: for example, really huge archives [2], [3] of scanned newspapers with heterogeneous content. Encyclopedia is also a good example of a document with heterogeneous content because we meet on its pages not only text, but a variety of content types including images, mathematical and chemical formulas, musical scores, technical drawings, chess notation, electronic circuits, etc., with varied geometric shapes.

In our work, we propose a toolkit in the form of a platform for processing documents containing heterogeneous content. We suppose also that any heterogeneous content is associated with the possibility of its presentation in a scripting language. The main features of such content 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 re-rendered from the scripts [1].

The problem of document image processing includes tasks that have already become almost classical, such as segmentation, noise filtering and image extraction from the background, determination of object boundaries, and pattern recognition.

The following describes one of the subtasks of the platform for document processing, namely segmentation, i.e., the process of selecting existing objects in the image by using existing software.

2. Development of a platform for solving the problem

Analysis of the work cycle for processing historical documents with a heterogeneous context showed that only a semi-automated workflow organization is possible. For implementation, a convergent technology for assembling complex software systems from ready-made modules a single platform is used; each of modules performs a small part of the task.

Let's consider the main ready-made systems used to develop the platform.

The main principle of our development is the platform virtualization. For this purpose, Docker[1] is used.

[1] https://www.docker.com/

The Docker system consists of a service that performs all the basic operations, two shells, the command line and the graphical ones, and the docker-compose extension.

Docker runs programs in an environment that is isolated from the operating system. It works on Linux, MacOS and Windows 10+. Docker *images* are quickly built and start. Each image, when launched, spawns its runtime instance called *container*.

An image is a collection of immutable *layers* that represent differences in the states of the filesystem. Images are immutable, and containers, in addition to the immutable layers of the image, contain a writable layer that reflects its changes. The *docker commit* command creates a new image from the container, adding its writable layer on top of the base immutable layers and making it a new immutable layer.

The developer is provided with a version tracking system similar to GitHub, with the ability to rollback.

A group of containers launched via docker-compose has access by default to the internal network that connects these containers. Each container can also connect to host networks either through port forwarding or directly. Finally, you can build an additional network from selected containers.

The containers' RAM is isolated from the host and from each other. Container disk storage can be modeled on host disks, on another machine, or in the cloud. You can also permit the container direct access to the host disk.

A library of ready-made images Docker Hub[2] is available.

The commercial system ABBYY FineReader Engine (FRE) is used to analyze the scanned document.

Image processing is performing with ImageMagick batch utility.

Another method used in platform development is Deep Learning ([4], [5]) whose implementation templates are widely available. This approach seems very suitable for our goals.

The language of implementation is Python. Its rich libraries provide a lot of ready-made solutions for subtask of the implemented platform.

[2] https://www.docker.com/products/docker-hub

3. Semi-automated workflow for recognition of heterogeneous documents

Despite a lot of achievements, automated recognition of the heterogeneous content remains a difficult problem. The problem is to maximize the support of semi-automated work.

To solve this problem, we propose a platform for recognition of heterogeneous documents, which uses existing and newly developed programs, and can perform all stages of the process.

The platform is created to integrate all existing software to maximize the degree of recognition automation. The recognition of heterogeneous documents involves many processes. Some may be performed automatically (green in Figure 1) using specialized software. Some processes need slight manual intervention or manual control. If the specialized software does not exist, the processing is executed manually (red in Figure 1) under the general purpose software. The integrating platform should facilitate the manual operations if they are necessary.

In Figure 1, after the preparatory stage the scan of the document is processed by a Python script developed by us, which analyzes the image. As a result of partitioning into regions, we obtain segment files with a uniform context (Scan segments) and page maps.

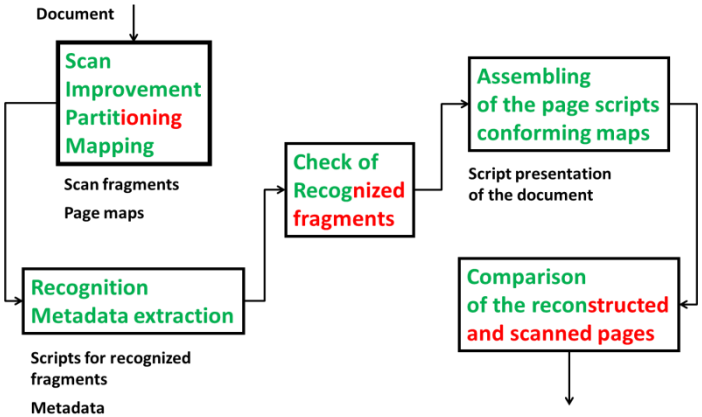


Figure 1. Structure of platform for recognition of heterogeneous documents

Using software for recognition and metadata extraction, we get scripts for the recognized segments and metadata.

The next stage is the check that can be automatic, or semi-automatic, or manual, depending on the content.

Finally, the assembly of the page scripts conforming to maps follows.

The result is a script presentation of pages of the document that can be used as a logical representation both for storage and for further processing.

Therefore, we can group all involved processes as follows:

- Automated: scan; segment recognition according to types of segments; assembly of script presentation of pages; metadata integration; reconstruction of page images from the script; automated verification.
- Semi-automated: image quality improvement; page layout analysis; task distribution for manual verification.
- Manual: human verification and manual correction.

Manual verification will be performed by experts in the corresponding areas. It implies that the platform will be Web-based like Wikipedia or version control systems.

The coordination of this activity is also necessary. Web platform for recognition of heterogeneous documents could be implemented to integrate all used tools.

4. Subtask: analysis of a scanned document and cutting it into segments with the same content type

ABBYY FineReader Engine (FRE) includes a ready-made command line interface (FRE CLI) that performs the full recognition cycle for one page at a time. The result is returned in XML format and contains the coordinates of the page segments, their type (text, image, table, separator, etc.), and the recognized text for text segments. Thus, it is possible to cut the page into parts with homogeneous content.

To process several pages, the utility can be called in a loop from a command script in any suitable language, including Python. It is possible to use container technology to build a platform from separate scripts.

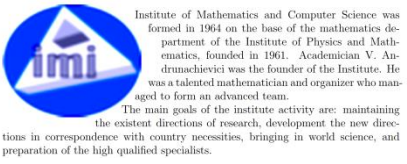
A Python program was developed and written to cut a scan of a document into segments with the same content type. The algorithm is as follows:

1. Using the FRE CLI utility, we analyze a page scan getting an XML file with the coordinates of the upper and lower corners of the segment rectangles and the segment type (text, picture, table, etc.).

2. A batch of scans of a multi-page document is processed in a cycle. Names of files with page images are set in the command line, with the ability to use regular expression elements (placeholders * and ?). For each image, a subdirectory is created with a name derived from the image name, into which the XML file and page segments are placed in a format that matches the page image format (Figure 2).

The Python script reads an XML file, selects the segment metadata (coordinates, segment type) and calls batch utility ImageMagick for image slicing. Separators are excluded from further processing.

3. After processing with FRE, each fragment on the page is described in an XML <block> element. The fragment's geometry is set by the <block> tag parameters: coordinates of the upper left and lower right corners of the minimal enclosing rectangle. A more complete description of the fragment's geometry is contained in a nested <region> element, which, in its turn, contains <rect> elements, each of which describes one rectangle. A fragment (“region”) consists of one rectangle or of several rectangles in more complex cases (Figure 3). For example, one line in a rectangle is such a region, if the analyzed fragment is text. If there are several such rectangles in the region, then the program provides a restructuring module to process such fragments. The module uses ImageMagick in a loop to compose a fragment from the constituent rectangles. The glued image fragment is written to a file for further processing.




```
<region>
<rect l="682" t="102" r="1884" b="159"/>
<rect l="683" t="159" r="1884" b="160"/>
<rect l="740" t="160" r="1884" b="463"/>
<rect l="682" t="463" r="1884" b="521"/>
<rect l="683" t="521" r="1884" b="523"/>
<rect l="623" t="523" r="1884" b="583"/>
<rect l="90" t="583" r="1884" b="762"/>
</region>
```

Figure 2. Image and set of coordinates for the text fragment produced by FRE CLI

The restructuring module allows one to solve the problem of reconstructing the geometric shape of a fragment. After executing the script, the following result is obtained.

For each page image, a subdirectory is created with the name generated from the image file name.

These subdirectories contain images of page segments recovered after restructuring. The subdirectory also contains two XML files. One of them is generated by the FRE CLI utility and describes the entire page. There are coordinates of all blocks, regions and rectangles, as well as the recognized text. The second XML file is generated in Python and contains a list of page segments with segment types and coordinates of the enclosing blocks.

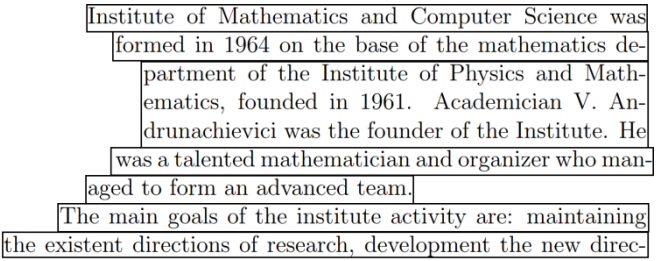


Figure 3. Fragment of irregular shape composed from several rectangles

Thus, a script developed in Python that uses FRE and the ImageMagick batch utility, generates files with scans of document segments and XML files with metadata for further processing by the platform, implementing a subset of the platform functionality.

4. Conclusions

Despite a lot of achievements, automated recognition of the heterogeneous content remains a difficult problem. We proposed a design of Web platform to maximize the support of semi-automated work of all used tools for recognition of heterogeneous documents.

In implementation, a convergent technology for assembling complex software systems from ready-made heterogeneous modules on a single platform is used. Each of modules performs a small part of the task using inside a container.

The program for partitioning and mapping of heterogeneous documents into homogeneous segments with different shapes was developed using ABBYY FineReader Engine, ImageMagick and Python. Scans with text, music, images, etc. were analyzed.

This is an example solution of segmentation and markup problem. The problem of complete classification for types of heterogeneous content isn't solved being the next stage of our work.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] A. Colesnicov, S. Cojocaru, and L. Malahov. *Recognition of heterogeneous documents: problems and challenges.* Proceedings of the 5th Conference on Mathematical Foundations of Informatics. 3-6 July 2019, Iasi, Romania, pp. 231-245, Iasi: „Alexandru Ion Cuza" University Publishers.

[2] Hui-Yin Wu and Pierre Kornprobst. *Multilayered Analysis of Newspaper Structure and Design.* [Research Report] RR-9281, UCA, Inria. 2019. hal-02177784

[3] Lee, B.C., Mears, J., Jakeway, E., Ferriter, M.M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., and Weld, D.S. *The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America,* 2020. ArXiv, abs/2005.01583.

[4] Simon Haykin. *"Neural Nemworкs: A Comprehensive Foundation"* Second Edition. [In Russian.] Саймон Хайкин. *«Нейронные сети: Полный курс»* Второе издание : Пер. с англ. М. – Издательский дом «Вильямс», 2006. 1104 с.

[5] Nikolenko S., Kadurin A., and Arkhanghelskaya E. *Deep Learning.* [In Russian.] Николенко С., Кадурин А., Архангельская Е. *Глубокое обучение.* – СПб.: Питер, 2018. – 480 с.

A. Colesnicov[1,2], S. Cojocaru[1,3], L. Malahov[1,4], L. Burtseva[1,5]

[1] V.A.Andrunachievici Institute of Mathematics and Computer Science
[2]E-mail: acolesnicov@gmx.com
[3]E-mail: svetlana.cojocaru@math.md
[4]E-mail: ludmila.malahov@math.md
[5]E-mail: luburtseva@gmail.com

On XML Standards to Present Heterogeneous Data and Documents

Alexandru Colesnicov, Ludmila Malahov,
Svetlana Cojocaru, Lyudmila Burtseva

Abstract

The article discusses XML presentation of heterogeneous data on cultural heritage in the form of Romanian documents in the Cyrillic script. An example of such presentation is standards for health care records. Currently a suitable framework for development of necessary presentation is Europeana Data Model. The necessary adaptation of presentation standards are considered.

Keywords: cultural heritage, digitization, heterogeneous data, XML-based presentation standards.

1 Introduction

Old documents subject to digitization often contain heterogeneous content: text, images, musical scores, mathematical formulas, etc. The recognition of these documents produces not only text outputs and images in electronic form, but script presentations of other fragments [1]. The recognition of heterogeneous content needs specific recognition agents for each type of page fragment and generates several resulting files.

Storing and searching the recognized heterogeneous documents is a special challenge also.

An example of successful solution of these problems exists in the domain of medical documentation. XML is used as top-level structure for document presentation due to its inherent extensibility.

For cultural heritage objects, the XML-based Europeana Data Model (EMD) is a suitable base to define their corresponding description [2].

The paper at first presents our approach to the digitization of heterogeneous documents. Section 3 describes examples of standardized XML-presentation of heterogeneous data: US standards for health care records, and EMD that is the modern approach oriented to cultural heritage. Section 4 sums up ways to configure our data structures in the EMD framework.

2 Digitization of heterogeneous documents

Heterogeneous documents, or documents with heterogeneous content, contain fragments that obey certain formal rules. Examples are mathematical and chemical formulas, chess notation and diagrams, electronic circuits, etc. The main properties of such content: 1) it is not a text in a natural language nor pure image; 2) there is a scripting language for its description; 3) a graphical presentation can be reproduced from the script. During digitization we should obtain the script presentation of all such fragments in the document. We described the process of heterogeneous document digitization in [3].

The digitization is performed in 11 stages. It uses and produces the following types of source, intermediate, and *final (7 types of 10)* data.

A Graphical document
B *Page image in electronic form*
C *Page map*
D *Page fragment*
E *Script equivalent of a page fragment*
F *Extracted metadata*
G *Script equivalent of a page*
H *Reconstructed page*
I Verification log
J Error report

All final data may be used in further work with the digitized docu-

ment; they and all their parts should be stored and available for search, and even for execution. This makes the structure of the digitized document unexpectedly sophisticated.

The usual and obvious solution for the presentation of the digitized heterogeneous document is to base this presentation on XML.

3 Examples of standard presentation of heterogeneous data

An important practical experience of data standardization is accumulated in health care, especially in the USA [4].

Standards used across health care organizations fall into four large groups: terminology standards, content standards, data exchange or transport standards, and privacy and security standards.

Less attention is paid to standardization of data store formats. Standards are applied starting from the API level. If a standard API call returns data in a standard format, the internal representation of these data is unimportant.

Non-textual and non-image information is kept and transmitted encoded. There are a lot of standard codes: codes for diagnoses, codes for medical procedures, codes for all kinds of health-related services, codes for dental treatment, codes for clinical information, codes for lab orders and results, codes for pharmacy products, and codes for clinical drugs.

As to images, Digital Imaging and Communications in Medicine (DICOM) is an international standard supported by all medical devices. The DICOM file has the extension .DCM and contains metadata plus from 0 to 7 images.

Health care data transmission uses XML and JSON formats.

Our second example is EDM [2], a model developed in the frame of Europeana project. The web-portal of unified access to growing digital cultural resources, Europeana, was launched in 2009 as a result of massive digitization of European cultural heritage in 2000s. Its basis is the metadata standard Europeana Data Model that was announced

in 2010. The main aim of the EDM is to provide a simple workflow for adding any local collection to Europeana portal. Thus, EDM elements set can be extended by each new provider, as he joins the Europeana information space. Besides the new elements introduced by Europeana, EDM has the set of elements re-used from other namespaces. It is supported by clear documentation and schema checking tools. Thanks to such support, EDM has become the XML-based standard for cultural objects metadata representation.

EDM replaces and widely extends the first model, which was called European Semantic Elements (ESE). More exactly, ESE was "the lowest common denominator" of semantics used in different cultural heritage sectors, like museums, libraries, archives, and audiovisual collections. EDM reverses this approach and covers all community standards, for example, LIDO for museums, EAD for archives, and METS for libraries [5, p. 4–5].

Today, Europeana joins over 58 million cultural heritage items from around 4,000 institutions. The ascent to BigData level creates problems common for such huge collections. To facilitate Europeana management the portal was divided into two ones:

- **Europeana,** which deals with ready collections only, provides their store, search and metadata retrieving;
- **EuropenaPro,** which deals with technical infrastructure, develops and maintains technical solutions for showcasing, sharing and using digital cultural heritage. All products developed in the frame of EuropenaPro are free and mostly open sources. EuropenaPro has own folder github.com/europeana, where open sources solutions can be obtained.

Having a long history and significant results, Europeana does not intend to stop its development. In 2020 a new Europeana strategy (2020-2025) was issued [6]. It was declared, that in its further development Europeana will focus on supporting the digital transformation of European cultural heritage sector. The tasks and priorities for both collection management and technical solutions development were revealed.

4 Data structure configuration in the EMD framework

We see that the internal representation of the document is not standardized. Externally, only the metadata set and correspondence to data exchange standards are important.

Our project supposes the development of heterogeneous document representation based on XML. The metadata should correspond to the EDM that guarantees extensibility and flexibility.

Each digitized document should be kept as an archive containing all its outputs (texts, scripts, and images). The Open Office DOCX format gives us an example: it is a ZIP archive containing XML files with texts included inside XML and images. One of XML files describes the document structure and lists other files. XML files also contain metadata related to the document, for example, the author name.

We could adopt this structure. Scripts for specific non-textual content could be kept inside XML like usual texts but marked as to be rendered by specific agents that makes them similar to images.

Each type of content and each type of output have its specific set of metadata. For example, page image should be accompanied by its page number.

5 Conclusion

The study illustrates the possibility of adapting the existing models in order to solve the problem of heterogeneous data and documents presentation. Taking into account the fact that we operate with heterogeneous elements within the printed texts, the metadata spectrum can be reduced and connected to the types of components, specific to these texts. Keeping the processed document as an archive, containing all the final elements listed in Section 2, seems to be a plausible solution.

Acknowledgments. This work was prepared as part of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

References

[1] A. Colesnicov, L. Malahov, S. Cojocaru, and L. Burtseva. *Semi-automated workow for recognition of printed documents with heterogeneous content.* Computer Science Journal of Moldova, vol. 28, no. 3(84), 2020, pp. 223–240.

[2] Available at: `http://pro.europeana.eu/edm-documentation`

[3] A. Colesnicov, S. Cojocaru, and L. Malahov. *Recognition of heterogeneous documents: problems and challenges.* In: Proceedings of the 5th Conference on Mathematical Foundations of Informatics, 3–6 July 2019, Iași, România, pp. 231–245. – Iași: Editura Universității "Alexandru Ion Cuza", 2019. – ISBN: 978–606–714–481–9.

[4] *Data Standards in Healthcare: Codes, Documents, and Exchange Formats.* October 23, 2020. Available at: `https://www.altexsoft.com/blog/data-standards-healthcare/`

[5] *EDM Primer.* Available at: `https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf`

[6] *Strategy (2020–2025). Empowering digital change.* DOI: 10.2759/524581. Available at: `https://pro.europeana.eu/files/Europeana_Professional/Publications/EU2020StrategyDigital_May2020.pdf`

Alexandru Colesnicov[1,2], Ludmila Malahov[1,3], Svetlana Cojocaru[1,4], Lyudmila Burtseva[1,5]

[1] "V. Andrunachievici" Institute of Mathematics and Computer Sc e, Chisinau, Republic of Moldova
[2] E–mail: `acolesnicov@gmx.com`
[3] E–mail: `ludmila.malahov@math.md`
[4] E–mail: `svetlana.cojocaru@math.md`
[5] E–mail: `luburtseva@gmail.com`

Development of a platform for heterogeneous document recognition using convergent technology

Alexandru Colesnicov, Ludmila Malahov,
Svetlana Cojocaru, Lyudmila Burtseva, Tudor Bumbu

Abstract

The paper describes architecture of a Web platform for recognition of heterogeneous documents. The platform provides comfortable working environment for users. The development of the platform uses the convergent technology that permits to integrate external applications easily, and maximally simplifies communication of the platform components.

Keywords: computer science, heterogeneous document recognition, framework architecture, integration of external applications.

1 Introduction

While digitizing texts, we often encounter documents with heterogeneous content. Each content type is recognized with its corresponding program. We need therefore a platform that combines all the necessary tools [1].

The proposed platform for digitization of heterogeneous documents supports all processing steps starting from electronic copies (scanned images). The platform supports image preprocessing, cutting of pages into homogeneous fragments, recognition of fragments, post-recognition processing, assembling of recognition results, and page reconstruction.

We discuss below architecture of the platform, the technology used at its development, and steps of digitization.

2 Platform architecture and development

The platform is a Web application consisting from user interface on the client side (frontend), and services implementation on the server side (backend). In our case, some functions are implemented on the client side that is described below.

The frontend is developing in Javascript using available standard packages and libraries (React, and more). Some necessary functionality is implemented in Javascript, for example, elementary image preprocessing. We use the corresponding packages, and implement some minor operations in the frontend instead of backend.

The backend is developing in Python using its rich libraries, and also calls external applications. The latter may be installed on the server, for example, ABBYY FineReader engine, and ImageMagick, or residing in the Web.

The convergent technology of the development [2] supposes smooth integration of the ready-made external applications. If the application has its documented API, this API is used. If the application doesn't have any API, it is executed in an isolated environment (sandbox). In both cases, exchange of data with the platform is performed through files.

The development is performed using Github version manager.

3 Digitization step-by-step

Steps of digitization supported by the platform are: uploading images and/or PDF files; image preprocessing; image fragmentation to parts with homogeneous content; recognition of fragments; post-recognition processing ot the results; assembling results; saving and downloading the results; restoring pages.

Comparison of restored and original pages is made manually but can be implemented lately.

Step 1. Uploading files. One or more files can be processed in a single digitization cycle. The following file types are supported:

PNG, JPEG, GIF, TIFF, and PDF. The total size and size of each file are restricted. File selection is performed through dialog, or by drag-and-drop.

2. Image preprocessing. This step is performed to obtain images with the quality suitable for recognition. Several preprocessing engines are available through submenu: Open CV, FineReader, ScanTailor, Gimp, ImageMagick. Open CV is available in Javascript and is executed on the client side. After selecting the engine, its specific options are offered.

Step 3. Fragmentation. This step permits to select image areas with homogeneous content, and detect the content type. The process is semi-automated. We use ABBYY FineReader engine on the server side to fragment images and preliminary detect fragment types. The fragmentation proposed by the program may be corrected manually. For the moment, the previewed types of content are: text, musical scores, mathematical formulae, chemical formulae and structures, chess diagrams. All unclassified content is marked as images. The platform is open for extensions; other content types may be added.

4. Recognition. Each type of the document content is recognized by its specific engine (FineReader, Mathpix, etc.) working on the server side, and the results are textual or script presentations of the content. It may be text in natural language for textual content, LaTeX script for mathematics, MusicXML for scores, etc.

5. Post-recognition processing. The obtained recognition results can be immediately checked and corrected. In some cases, transliteration of the text may be necessary, for example, from the old Romanian Cyrillic to the modern Latin alphabet. For the latter task, we implemented a dedicated application AAconv. The result of transliteration can be corrected manually.

6. Assembling recognition results. All original page images are collected in a PDF file. This is made even for pages of original PDF file because that file may be blocked for changes. The PDF standard permits to integrate inside a PDF other files as attachments. All results of recognition are attached for each page. Equally, page maps are

attached that contain coordinates of each page fragment.

7. Saving and downloading the results. The assembled file that integrates original images with recognition results and page maps is saved on the server side and can be downloaded by the user.

8. Restoring pages. Scripts permit restore graphical presentation of recognized fragments using applications that correspond to their content types.

4 Conclusion

The developing platform integrates in a unified comfortable environment all tools that are usually used at digitization of documents with heterogeneous content.

Acknowledgments. This work was prepared as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] A. Colesnicov, L. Malahov, S. Cojocaru, and L. Burtseva. *Semi-automated workow for recognition of printed documents with heterogeneous content.* Computer Science Journal of Moldova, vol. 28, Nr. 3, 2020, pp. 223–240.

[2] A. Colesnicov, L. Malahov, S. Cojocaru, and L. Burtseva. *On convergent technology in development of information systems for processing of documents with heterogeneous content.* In: Proceedings of the Workshop on Intelligent Information Systems, 04–05 December, 2020, Chișinău, Republic of Moldova, pp. 61–68, ISBN 978–9975–68–415–6.

Alexandru Colesnicov, Ludmila Malahov,
Svetlana Cojocaru, Lyudmila Burtseva, Tudor Bumbu

Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mails: alexandru.colesnicov@math.md, ludmila.malahov@math.md, svetlana.cojocaru@math.md, liudmila.burteva@math.md, tudor.bumbu@math.md

Researching and valorizing the lexicon of the Romanian Language in a general Romanian context

Olesea Caftanatov, Ludmila Malahov

Abstract

The aim of our research is to initiate the creation of lexicographic corpus consisting of dictionaries developed over the years by specialists from Institute of Philology. Additionally, we developed the project's website "LogosPlus", which includes online library, online dictionary, blog and other related resources.

Keywords: lexical borrowing, academic dictionaries, digitization, virtual library.

1. Introduction

Speaking about the active processes that occur in the lexicon of the current Romanian language, we mention the extension and restriction of the senses, metaphorizing, metonymizing, forming new words and meanings, and borrowings. *"The Romanian language is very open to novelty"*, thus, one of the priorities of the project is the creation of a bank of recent lexical acquisitions (based on the monitoring of the current print, electronic and audiovisual press in Moldova), in order to supplement the dictionaries with recent words and meanings.

At the same time, in today's computerized society, each language needs technological products to connect it to the international circuit of communicative techniques and tools. This is why, we developed a blog that will contain the post with results of our research (more about this is presented in Section 2).

Additionally, we intend to digitize dialectal texts and transliterate them in contemporaneous Romanian language alphabets. For this purpose, we digitize and transliterate the book "Тексте Диалектале" [1] (more about this see in Section 3).

2. Blog "Cultivating Romanian Language"

Due to the new technologies, the best solution for disseminating the results of our research is by creating digital text. Thus, we developed "Cultivating Romanian Language" as a blog, where each post contains the result of our investigation [2]. To date we created 36 posts that can be classified in seven main threads, such as lexical rules, grammar, morphology, orthography, semantics, syntax and stylistic, see Figure 1. Additionally, the content can be freely accessed and shared on other social media platforms such as Facebook, Twitter, LinkedIn etc.

Figure 1. The Blog's "Cultivating Romanian Language" interface

3. Digitization and transliteration

One of our project's goals is the creation of a lexicographic corpus consisting of dictionaries developed over the years by specialists from the Institute of Philology, especially from the sector of lexicology and

lexicography, which would complete the essential Romanian lexicographic corpus (100 dictionaries from the DLR bibliography).

Moreover, an important task regarding the creation of lexicographic corpus was to digitize and transliterate books from Soviet period [1, 3-5].

The most interesting in sense of technology development was the book "Тексте Диалектале" ("Dialectal text"), because in order to recognize phonetic features we created a special alphabet, that consists of Cyrillic alphabet, diacritics and conventional text.

Below we will describe few steps regarding whole digitization process. First, the digitization has several stages: scanning, post-processing of the scanned resources, preparation for recognition, optical character recognition (OCR), automatic and manual validation.

Regarding the first step, we carried out scanning with an accuracy of 600 dpi and the resulting files were placed in the Vladimir Andrunachievici Institute of Mathematics and Computer Science cloud. For post-processing step we used Scan Tailor application. It is an interactive post-processing tool for scanned pages. It performs operations, such as page splitting, deskewing, adding/ removing borders, binarising, cleaning, among others features [6].

Before OCR step, we prepared templates with special alphabet and extended existed dictionary. For recognition step, there was used ABBYY FineReader v.14, which was tuned to a specific phonetic alphabet of the Moldovan language (Romanian in Cyrillic). To improve recognition, we used a custom dictionary from the previously correctly recognized words from the book.

Automation validation was carried out in the editing mode with a hint from constructed dictionary. After that, only about 5-10% errors were corrected manually. To date, the recognition of volume 1 of the book has been completed. The texts are saved in the cloud: https://cloud.math.md/.

After recognition, the important task is transliteration into Latin alphabet. It should be noted, that there are dissimilarity of phonetic records of texts in Cyrillic and Latin, since in the phonetic alphabet of the Cyrillic alphabet there are Latin letters, Greek, thus, transmitting their non-standard pronunciation. Therefore, for transliteration we created 273 rules. Some of the rules for overlapping symbols can be seen in Table 1.

There was a problem of their classification, which was solved by successive partitions and recursion on test examples. However, for this process, manual validation is necessary in order to get good results.

Table 1. Examples of rules for overlapping symbols

а о	→ a/o	э а	→ ă/a	́э Р	→ ắ/P	т" к"	→ ä/W
а ы	→ a/î	э е	→ ă/e	¼ л	→ ¼/l	е и	→ e/i
а э	→ a/ă	́э а	→ ắ/a	¡ н	→ ¡/n	é и	→ é/i
и е	→ i/e	ы и	→ î/i	ы э	→ î/ă	h х	→ M/h

4. Virtual library

Another object of our research is creation of a virtual library (see Figure 2.A). This library will contain two data bases, one - for articles (see Figure 2.B) and another one - for monographs (see Figure 2.C).

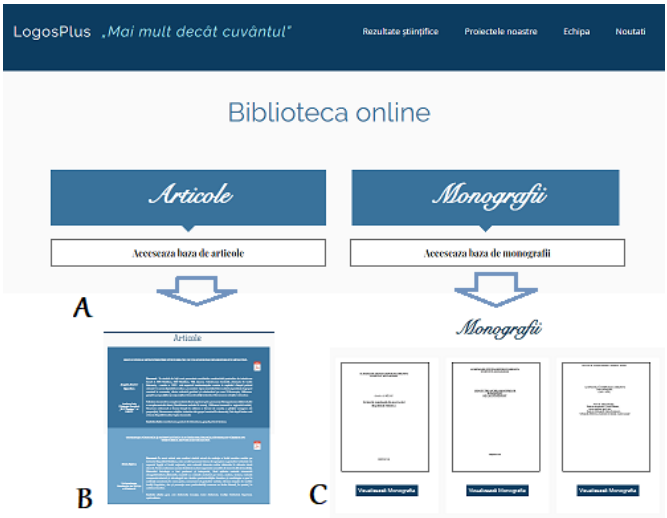


Figure 2. The Virtual Library Interfaces

5. Conclusion and future work

For our project, we recognize the book "Тексте Диалектале" vol. 1. Moreover, we elaborated 273 rules for transliteration. For future work we will continue to research and valorize the lexicon of the Romanian Language by digitizing and transliterating other books. The recognized books would be placed in our virtual library form our official project web site. In addition, we created a bank with lexical acquisition and will continue to extend it. The research results will be published on: "Cultivating Romanian Language" blog.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" and the project "Scientific valorization of the national linguistic heritage in the European integration context".

References

[1] V. Melnic, V. Stati, and P. Udler. *Texte dialectale*, vol 1. Chișinău, 1969, pp. 12-238.

[2] Project's web site Logos Plus. www.logosplus.org

[3] G. Buciușcanu. *Gramaticii limbii moldovenești*. Editura de Stat a Moldovei, Balta (1926).

[4] L. Madan. *Gramaticî moldovneascî* (1930). Partea I. Fonetica și Morfologhia. Tirișpolea, 1930.

[5] I. Cușmaunsa. *Gramatica limbii moldovenești*. Manual pentru școala necomplectă și mijlocie. Tiraspol, 1939.

[6] Scan Tailor official web site. https://scantailor.org/

Olesea Caftanatov[1], Ludmila Malahov[2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: olesea.caftanatov@math..md

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: ludmila.malahov@math.md

Towards a Font Classification Model for Romanian Cyrillic Documents

Tudor Bumbu

Abstract

This paper presents a solution on how to classify the fonts in the 17th century Romanian Cyrillic documents. This solution is based on a mix of unsupervised and supervised machine learning technics. The unsupervised process is the application of K-Means method to create the dataset with the fonts characters and their labels, whilst the supervised process is to train two different architectures of neural networks to classify these characters.

Keywords: old documents, OCR, font classification, neural networks, Romanian Cyrillic.

1 Introduction

Some Romanian Cyrillic documents of the 17th century require particular models at optical character recognition (OCR) because some printing houses used different character printing styles (fonts) than others [1].

The problem of identifying and classifying the font in a document printed in the 17th century can be formulated as follows: *Given a document X from 17th century printed in Cyrillic Romanian and a set N of OCR models trained on documents of the 17th and 18th century, choose the most appropriate model for X.*

A trivial solution is to recognize a sample (a page snippet) from document X using all models in N and basing on the results, to choose the model that gives the highest accuracy (best result). The time complexity of this solution does not fit our needs, as we have to load all OCR models, recognize the snippet and measure the accuracy for each model separately. Model upload time and sample recognition can

exceed 2 minutes depending on page size, and if we have 10 different models, we have to wait for approx. 20 minutes to find the the most appropriate model each time we want to recognize a document of this kind.

The proposed solution is to train a neural network with samples from several Romanian documents printed in the 17th century at different printing houses. A neural network is trained with a dataset consisting of tuples of *image character* and its *class* in order to be able to further classify a new sample.

In the next section we describe the selected document samples aiming at creating the dataset.

2 Dataset resources

The Romanian Cyrillic alphabet was used at printing houses in regions as *Iași, Bucharest, Târgoviște, Belgrade (Alba Iulia), Uniev (Cernăuți), Sas Sebeș, Snagov, Buzău.* Each of this region had at least one printing house in the 17th century.

The data set is created from 10 scanned books, selected from the digital library of Romania (http://digitool.bibnat.ro). In the selected books, two distinct sets of characters were observed. Therefore, the books were divided in two classes depending on their font style: one class of the books consisting of 13 pages was included in the set A, and the other class with 9 pages was included in the set B. Figure 1 shows two samples from each set A and B. Two main letters that differ in both samples are: *m* and *з* (*t* and *z* in Latin).

In the next section, we describe the main tasks in the process of creating the dataset: segmentation of text areas in the pages from A and B; detection of individual characters in text blocks; clustering the characters and forming the training and testing data set.

3 Creating the dataset

In the subsections below we describe tools for segmenting the regions of text (text blocks) in the selected pages, a method for identifying the individual characters in a text block and approaches to group characters in similar groups in order to create a better dataset.

Figure 1. A sample from set A (on the left) and one sample from B

3.1 Segmentation of text blocks

From the pages prepared for the dataset we extracted fragments of text using *Detectron2* [2] segmentation tool trained on the *PrimaLayout* dataset [3].

In *PrimaLayout* dataset, the text portions are labeled with the label "TextRegion". We have extracted more than one block of text from every single page based on this label. For this reason, two blocks of text may contain the same characters, and similar examples of training and testing may appear in our dataset.

After segmentation, the text blocks were cut and placed in a list of text blocks. On average, 4 blocks of text were obtained for each of the 22 selected pages. These fragments are saved as images.

In the next subsection we identify and extract the characters from text blocks. Character set consists of *letters, punctuation marks, accents, outlines of tables, stain pixels,* or *page noise.*

3.2 Detecting individual characters of text blocks

We need to make sure that each image, independently of its source, is processed in such a way that the algorithm used to detect the letters can find as many letters as possible. So, we converted all images to black and white. As a result, the images consisted only of black and white pixels. We optimized the image for letter de-

tection using the *findCountours()* method within openCV library (https://docs.opencv.org/4.5.3/). We mapped the contour delimitation boxes to the original image to see what was actually detected. The result of this processing step is shown in Figure 2.

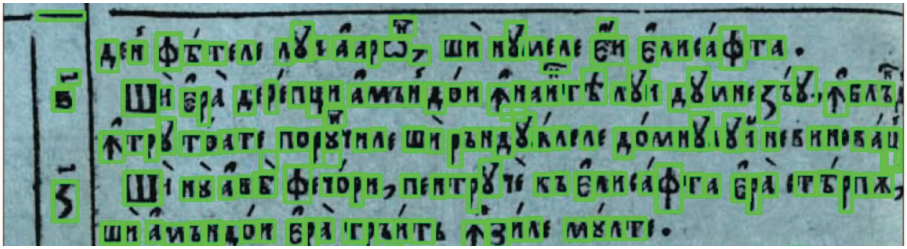


Figure 2. The contours of the characters in block of text from set *A*

The characters inside boxes (Figure 2) were cut and saved in two folders, for each class of fonts separately. From the first 10 pages of set *A* we obtained 17,155 characters, and from 9 pages of set *B* – 8,799 characters. Among the extracted characters, there are also elements of noise in the image – spots, accents without the basic character (letter), tilde, etc.

In the next subsection we try to remove unnecessary characters and keep only the letters. For this task we use *K-Means* clustering model implemented in *Scikit-learn* library (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html).

3.3 Clustering the characters

We had to organize the characters extracted from text blocks in such a way that these letters become useful to train a neural network. So, we removed images that do not contain letters; grouped all the remaining images by clustering (this means that all the letters *"ц"* form a cluster, all the letters *"з"* form another cluster, and so on).

Before clustering, we had to make sure that all the images are of the same size. Each extracted character has the size of its boundary box, which varies widely. So, we resized all the images to the size of *50 by 50 pixels*.

K-Means clustering method was applied because it is simple and fast. The only thing we had to provide is the number of clusters. For a perfect result we have to end up with the exact number of letters in the Romanian Cyrillic alphabet – up to 47 letters considering the uppercase and lowercase. Each of them can appear as a lowercase or uppercase letter, which means that, in total, we expected 100 clusters (including punctuation marks).

After analyzing the clusters and deleting characters that are not letters and punctuation marks, we placed them in two bigger clusters: one – for set A and another cluster of characters – for set B. There are more than 10,000 characters in the set A, and set B contains 8,775 characters.

To form a well-organized dataset, we converted the dataset to *IDX format*. Since we train a neural network, we have to divide the images into two sets: a set of training and a set of testing data. For this purpose, we moved 1/3 of the former sets A and B to a test set. The output is 4 files: 2 files for image characters – for training and for testing, and another 2 files with the labels. The labels are the values *0* and *1*, where *1* is the class label for a character from set A, and *0* is the class label for a character from set B. After counting the examples in the dataset, we have: 21,212 *training examples* and 9,093 *test examples*. In Figure 3, we can see some examples from the training dataset.

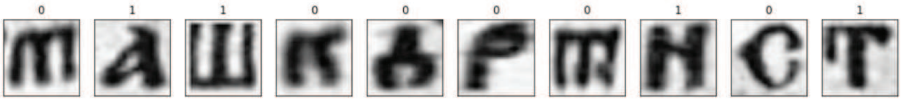


Figure 3. Samples from training dataset, where each image character has its class label – *0* or *1*

In the next section, we train a multilayer neural network and a convolutional neural network with the dataset prepared at this stage.

4 Neural network models

We trained two different neural network models to classify the characters in 2 classes of fonts. The first model is a multilayer perceptron (MLP), and the second one is a convolutional neural network.

The MLP model was implemented using Tensorflow and Keras (https://www.tensorflow.org/guide/keras). The input image matrix is reshaped into a vector of length x by y, where $x = 50$ and $y = 50$ – the dimensions of the image. Thus, we have an input layer of 2,500 neurons. We added a hidden layer with 128 neurons with the $ReLU$ activation function which is completely connected to the last layer. The last layer (output layer) contains a single neuron and a sigmoid function for its activation. At the training phase we set 300 epochs. The training phase lasted about 55 minutes without GPU.

The MLP model delivers an accuracy of 96.7%. Based on the confusion matrix (Figure 4), we computed the classification error – 3.3%.

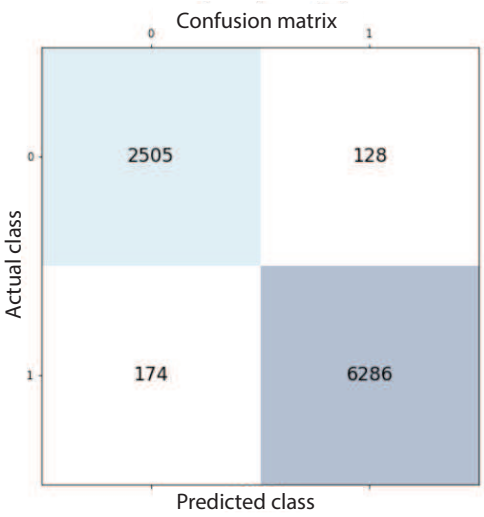


Figure 4. Confusion matrix of the MLP model based on the test dataset

The result we have seen after training the MLP model is pretty good, but there is a chance that we can get a better accuracy using an

architecture which is more specialized in image processing, namely the convolutional neural networks (CNN).

So, we have built a CNN model with 3 convolutional layers of 16, 32 and 64 neurons and one dense layer of 64 neurons. The output layer consists of 2 neurons as we use categorical cross-entropy as the loss function. The accuracy of our CNN model is 98.2% after training it through 100 epochs (see Figure 5). In this case, we used GPU power, and the training lasted about 11 minutes. We obtained a better result which can be the part of our final solution on classifying the font of a document X.

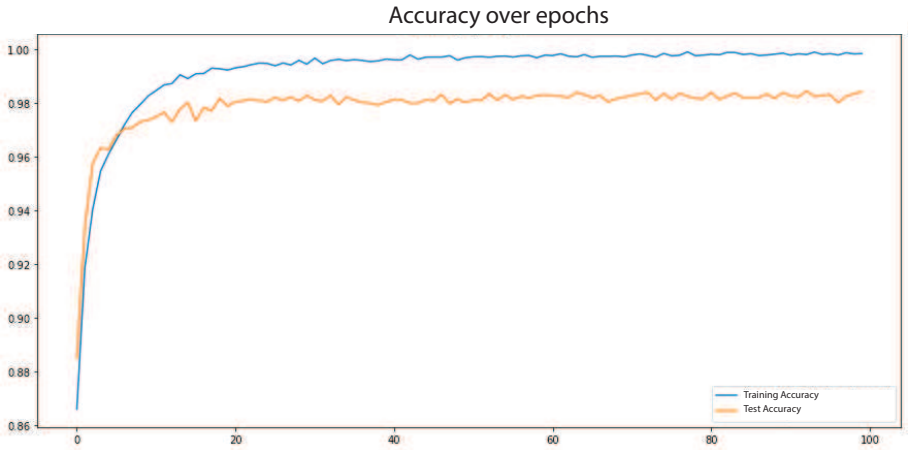


Figure 5. Accuracy chart for CNN model on training and test set

5 Conclusions

This paper is the extended and revised version of the conference paper [4] presented at WIIS 2021. In this paper we proposed a solution on how to classify the fonts in the Romanian Cyrillic documents of 17th century based on a mix of unsupervised and supervised machine learning technics. We used the unsupervised method K-Means to create the dataset with the font characters and their classes and two different neural network architectures MLP and CNN to classify these characters.

The obtained models perform well with the best accuracy of 98.2% by the CNN model. It can be deployed in our Model Selector tool [1] in order to identify the best OCR model to use on a particular Romanian Cyrillic document depending on its font style.

Acknowledgments. The research was supported by the project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge, and big data".

References

[1] T. Bumbu, S. Cojocaru, A. Colesnicov, L. Malahov, and Ș. Ungur, "User Interface to Access Old Romanian Documents," in *Proceedings of the 4th Conference of Mathematical Society of Moldova CMSM4'2017, June 25-July 2, 2017*, pp. 479–482.

[2] R. Girshick, I. Radosavovic, G. Gkioxari, P. Doll ar, and K. He, "Detectron," 2018. Available: https://github.com/facebookresearch/detectron.

[3] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Prima Layout," in *Proceedings of 14th International Conference on Document Analysis and Recognition (ICDAR), 2017, volume 1*, pp. 1404–1410.

[4] Tudor Bumbu, "On Classification of 17th Century Fonts using Neural Networks," in *Workshop on Intelligent Information Systems (WIIS2021)*, (Chisinau, Republic of Moldova), October 14-15, 2021, pp. 58–64.

Tudor Bumbu

Vladimir Andrunachievici Institute of Mathematics and Computer Science
5, Academiei street, Chisinau, Republic of Moldova, MD 2028
E–mail: tudor.bumbu@math.md

Backtracking algorithm for lexicon generation*

Constantin Ciubotaru

Abstract

This paper is dedicated to generating process of the Romanian Cyrillic lexicon used between 1967 and 1989. The rules for transliteration of words from the modern Romanian lexicon to their equivalents written in Cyrillic were established and argued.

A backtracking algorithm has been developed and implemented that generates the Cyrillic lexicon using the transliteration rules. This algorithm actually is a tool to facilitate the work of the expert. The work of the expert is reduced to checking the transliterated variants and changing the transliteration rules.

Keywords: lexicon, transliteration, backtracking algorithm, decyrillization, morpho-syntactic descriptions (MSD).

1 Introduction

The problem of digitizing and preserving the historical-linguistic heritage is a priority domain of the digital agenda for Europe. The digitization process requires solving a series of problems related to the recognition, editing, translation, and interpretation of printed texts. The solving of these problems for the Romanian historical-linguistic heritage faces difficulties and specific aspects: a large number of periods in the evolution of the language, a small volume of stored resources that are also scattered, a great diversity of alphabets.

The presence of a digitized Romanian Cyrillic lexicon will contribute to the regeneration, revitalization and preservation of the heritage related to this period. Various aspects of the problem have been exposed in [1]–[3].

*This work was supported by the project Nr. 20.80009.5007.22

The paper addresses the issues related to the digitization and transliteration of the historical-linguistic heritage printed in Cyrillic script during 1967–1989 on the territory of the Moldovan Soviet Socialist Republic (MSSR), in accordance with the linguistic norms of the modern Romanian language.

During that period the Moldovan Cyrillic alphabet (AlphaCYR) was used which actually represents the Russian alphabet without the letters "ё", "щ" and "ъ" and extended by adding the letter "ж" in 1967. Complete lack of resources in electronic format and presence of fragmentary grammatical descriptions that admit ambiguous interpretations represent the main difficulties specific to the period.

According to the dexonline definition, *transliteration* is the "transcription of a text from one alphabet to another, rendering the letters by their equivalents, regardless of the phonetic value of the signs" [4].

The process of transliterating Romanian words into their written equivalents with the characters of the AlphaCYR alphabet is called *cyrillization*. For instance, *"puiului"*⇒"пуюлуй", *"fiului"*⇒"фиулуй", *"cenușiu"*⇒"ченушиу", *"viermi"*⇒"вермь", *"vierii"*⇒"виерий".

The inverse procedure for cyrillization is called decyrillization, e.g. "пуюлуй"⇒*"puiului"*, "бьет"⇒*"biet"*, "боер"⇒*"boier"*, "пепт"⇒*"piept"*.

If the digitization of the text is relatively simple, the problem of recognizing the digitized text is quite complicated, especially considering the total lack of Romanian Cyrillic resources for that period. This paper extends the results presented in [5] and aims to develop a tool for generating the lexicon corresponding to that period (noted by LexCYR), starting from the lexicon of the modern Romanian language (noted by LexROM).

The general scheme of the Romanian Cyrillic lexicon generator is presented in Figure 1.

Figure 1. The general scheme of the Cyrillic lexicon generator

2 Selection of the modern Romanian lexicon

For the choice of the modern Romanian lexicon, the following three resources were examined:

1. ***Dexonline*** [4]. It contains over 900000 entries, with a convenient interface for online use. The dictionary structure is less adaptable for processing because it does not contain explicitly the inflected forms, does not contain morpho-syntactic descriptions (MSD), and includes both forms of spelling "î" from "i" and "â" from "a" ("fân"-"fîn", "pârî"-"pîrî").
2. ***The lexicon developed at the "Al.I.Cuza" University, Iași*** [6] with over 1000000 entries. The lexicon is well structured, contains MSD labels in accordance with the tagset proposed in the project MULTEXT-East [7]. But, as in dexonline, we find both spellings "î" / "â", also many proper names and words of foreign origin, to which the rules of transliteration cannot be applied.
3. ***Reusable linguistic resources*** developed at the Institute of Mathematics and Computer Science "Vladimir Andrunachievici" [8] with over 677000 entries, including inflected forms. The formalization (packaging) of resources is quite complicated, the morpho-syntactic descriptions are incomplete.

Finally, the lexicon developed at the "Al.I.Cuza" University (LexROM) was selected with minor modifications, as follows:

1. Proper nouns and words of foreign origin were removed;
2. All words were transliterated using the spelling "â" from "a" according to the provisions of the Romanian Academy. Duplications of spellings "î" / "â" were avoided by applying an algorithm specially developed for this purpose.

The problem of spelling "î"/"â" does not affect the cyrillization process, because in both cases there is the same result at transliteration: *"â"* ⇒ "ы", *"î"* ⇒ "ы". Difficulties arise in the decyrillization process: should we apply the rule "ы" ⇒ *"â"* or rule "ы" ⇒ *"î"*?

We denote by AlphaROM the Romanian language alphabet, and by LexROM(α) – all the words from LexROM that start with the letter α, $\alpha \in$ AlphaROM.

3 Used tools

To formalize the transliteration rules and program the lexicon processing algorithms there was selected the Common LISP functional programming language [9],[10].

The Notepad++ editor was used for word processing [11], which offers advanced editing capabilities, such as:

- select text both horizontally and vertically,
- store search results in separate files,
- mark lines and operations with these lines,
- allow the use of regular expressions,
- support UTF-8 encoding for Romanian letters with diacritics and Russian, for example: Ă, ț, Â, â, Ș, Э, ц, Ы, Ш, ӂ,
- rich set of plugins: exporting files in various formats (RTF, HTML), the ability to launch applications (files with the extension .exe), sorting and comparing files, etc.

4 Backtracking method

The backtracking method proposes to build the solution(s) of a problem incrementally by applying iterative and/or recursive algorithms. It is assumed that there is a finite set of candidates for solutions and some internal criteria for verifying candidates. The method can be applied to generate the lexicon, as all the necessary conditions are met:

- the modern LexROM lexicon is given,
- sets of rules for transliteration are defined,
- there is a finite set of intermediate transliterated words that represent candidates for solutions,
- there are internal criteria for verifying the variants: the order of application of the rules, context-sensitive dependencies, prefixing and suffixing, the involvement of the expert,
- the set of all solutions meets the LexCYR lexicon,
- iterative and recursive algorithms are applied.

5 Algorithm of switching to the spelling "â" from "a"

The transition to the spelling "â" from "a" also will be done by transliteration. According to the provisions of the Romanian Academy, the letter "î" will always be written at the beginning and end of the word ("început", "înger", "în", "întoarce", "a coborî", "a urî"). Inside the word, it is usually written "â" ("cuvânt", "a mârâi", "român", "fân"). There are, however, a few exceptions to this rule. Words formed by prefixing words that begin with the letter "î" will keep this "î" inside. For example, "neîmpăcat", "neîngrijit", "preîntâmpinat", "dezîntors", "reînarma". The same rule will be applied to compound words: "bineînțeles", "semiînchis", "altîncotro". There are also a few exceptions, for example, the word "altînghie" will be transliterated as "altânghie", because it is not a compound word, this is the name of a flower, also called "lady's slipper". On the other hand, the word "capîntortură" (the name of a bird) will be transliterated, together with its derivatives, as "capîntortură". It is taken into account that the word comes from "cap întors" ("turned head"). The specificity of the LexROM lexicon will also be taken into account, that includes, along with the lemma words and inflected forms, phrases and word combinations, which can be spelled with "î" from "i". These words inside the construction are separated by "~". For example, "pe~înserate", "de~jur~împrejur" etc. All words w that contain at least one letter "î" can be represented as $w = w_0 \bullet$ "î" $\bullet w_1 \bullet$ "î" $\bullet \ldots \bullet w_{n-1} \bullet$ "î" $\bullet w_n$. If the word starts with "î", then $w_0 =$ "". We will mark by "" the empty string. For words ending with "î" we will have $w_n =$ "". Thus, for the letter "î" we get "î"=$w_0 \bullet$ "î" $\bullet w_1$, $w_0 = w_1 =$ "". For the word "coborî" we obtain: "coborî"=$w_0 \bullet$ "î" $\bullet w_1$, with w_0="cobor", $w_1 =$ "". For the combination of words w="din~cînd~în~cînd" we have: $w = w_0 \bullet$"î"$\bullet w_1 \bullet$"î"$\bullet w_2 \bullet$ "î" $\bullet w_3$="din~c" $\bullet$"î"$\bullet$"nd~" $\bullet$ "î"$\bullet$"n~c" $\bullet$"î"$\bullet$"nd". Note that w_1="nd~" ends with "~", which means that the next word will start with "î", analogous to the prefix situation. As a result of the conversion we get "din~când~în~când".

Performing a statistical analysis of the LexROM lexicon leads to selection of the set of all prefixes that can be inserted in front of words

that start with the letter "î". This set is denoted by PREFIXES.

Algorithm of switching to the spelling "a" from "a"

0. ***Start***

1. The lexicon of the modern Romanian language LexROM is given.
* We will modify this lexicon by substituting all words with their written equivalents with "â" from "a" applying the transliteration method *\

2. We modify the LexROM by applying transliteration rules for exceptional situations. For example, "altîngie" ⇒ "altângie" (in other cases "alt" will be a prefix).

3. We build the set of prefixes that can be placed in front of words which start with the letter "î". PREFIXES={"alt" "arhi" "auto" "bine" "bio" "de" "dez" "din" "ex" "ne" "nemai" "ori" "piți" "pre" "prea" "pro" "re" "semi" "sub" "subt" "super" "supra" "tele"}.

4. ***loop for all*** $w \in$ LexROM ***do***

4.1. ***if*** w does not contain "î" ***then return***(w).

4.2. We represent $w = w_0 \bullet$ "î" $\bullet\, w_1 \bullet$ "î" $\bullet \ldots \bullet\, w_{n-1} \bullet$ "î" $\bullet\, w_n$, where $w_0, w_1, \ldots, w_n$ are words which do not contain "î", $n \geq 1$.

4.3. ***if*** $(w_0 =$ "") ***or*** $(w_0 \in$ PREFIXES) ***or*** $(w_0 = w_0' \bullet$"~") ***or*** $(w_1 =$ "") ***then*** $w_r := w_0 \bullet$"î" ***else*** $w_r := w_0 \bullet$"â".

4.4. ***loop for*** i ***from*** 1 ***to*** $(n-1)$ ***do***

4.4.1. $w_r := w_r \bullet w_i$

4.4.2. ***if*** $(w_{i+1} =$ "") ***or*** $(w_i = w_i' \bullet$"~") ***then*** $w_r := w_r \bullet$"î" ***else*** $w_r := w_r \bullet$"â".

4.5. ***end loop***

4.6. ***return***$(w_r \bullet w_n)$

5. ***end loop***

6. ***Stop***

6 The structure of the lexicons

The LexROM lexicon is represented as a list in Common LISP, each element of the list being composed of three components: (word, MSD-label, word-lemma). For each element of the LexCYR lexicon, the fourth component – the cyrillized word (Figura 2) – is included.

(ghiocei "Ncmprn" "ghiocel")	(гиочей "ghiocei" "Ncmprn" "ghiocel")
(ridic "Vmsp1s" "ridica")	(ридик "ridic" "Vmsp1s" "ridica")
(ridica "Vmn" "ridica")	(ридика "ridica" "Vmn" "ridica")
(ridicam "Vmii1p" "ridica")	(ридикам "ridicam" "Vmii1p" "ridica")
(ridicăm "Vmsp1p" "ridica")	(ридикэм "ridicăm" "Vmsp1p" "ridica")
(ridicare "Ncfsrn" "ridicare")	(ридикаре "ridicare" "Ncfsrn" "ridicare")
(ridicat "Ncmson" "ridicat")	(ридикат "ridicat" "Ncmson" "ridicat")
(ridicat "Afpmson" "ridicat")	(ридикат "ridicat" "Afpmson" "ridicat")
(ridicat "Vmp" "ridica")	(ридикат "ridicat" "Vmp" "ridica")
(ridicat "Rg" "ridicat")	(ридикат "ridicat" "Rg" "ridicat")
(ridicatele "Ncfpry" "ridicat")	(ридикателе "ridicatele" "Ncfpry" "ridicat")
(ridicatule "Ncmsvy" "ridicat")	(ридикатуле "ridicatule" "Ncmsvy" "ridicat")
(a) LexROM structure	(b) LexCYR structure

Figure 2. The lexicons structure

The MSD label is a set of characteristics of the word viewed as part of speech. The label represents a sequence of symbols, the first symbol specifying the part of speech (for example, N - noun, V - verb, A - adjective, Rg - adverb, etc). The rest of the symbols will specify the morphological characteristics of the word, such as number, gender, person, time, case, mode, etc. The scheme of the MSD label for the noun is shown in Figure 3.

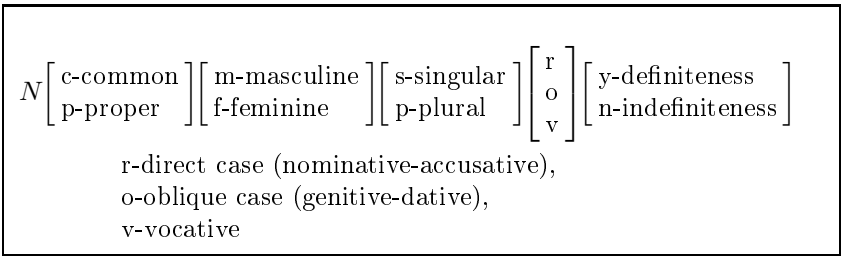


Figure 3. The MSD label structure for noun

The following algorithm based on the backtracking strategy is proposed for generating the Cyrillic lexicon LexCYR. Sets of transliteration rules are defined, cyrillization and decyrillization algorithms are constructed. The cyrillization algorithm is applied on the Romanian lexicon LexROM. A variant of the LexCYR lexicon will be obtained,

which can be subjected to decyrillization, thus a new variant for the Romanian lexicon is obtained. The ideal situation would be to match these two lexicons. If inconsistencies occur, the expert intervenes, who can change the rules of cyrillization\decyrillization, can repeat the whole process or can intervene with corrections on the constructed Cyrillic lexicon.

7 Cyrillization

Unlike the problem of digitizing and recognizing printed text, which is solved relatively simply, the problem of cyrillization is more difficult. To solve this problem we will apply the transliteration method. By definition, the transliteration process consists in the consecutive application of a set of substitutions (rewriting rules). For example, *brad* ⇒ б*rad* ⇒ бр*ad* ⇒ бр*а*д ⇒ брад. Here the following rules have been applied consecutively "*b*" ⇒ "б", "*r*" ⇒ "р", "*d*" ⇒ "д", "*a*" ⇒ "а". We will call these rules general rules. For them the order of application is irrelevant.

For the letter "*i*" we have the general transliteration rule "*i*" ⇒ "и", but the following rules are also possible: "*i*" ⇒ "й" and "*i*" ⇒ "ь". Examples: *fuior* ⇒ фуйор, *fior* ⇒ фиор, *miere* ⇒ мьере.

In other cases the rules may be more complicated. For example, two rules can be applied to the letter "*g*": "*g*" ⇒ "г", "*g*" ⇒ "ѓ" – *gigant* ⇒ ѓигант. Here comes the context-sensitive rule that requires substitutions: "*gi*" ⇒ "ѓи", "*ge*" ⇒ "ѓе", "*ghi*"⇒ "ги", "*ghe*" ⇒ "ге".

Thus, it becomes obvious that these rules must be applied before applying the general rule "*g*" ⇒ "г". Moreover, substitutions "*giu*" ⇒ "ѓиу", "*giu*" ⇒ "ѓю" are also possible. For example, *giulgiu*⇒ ѓюлѓиу, *giugiuli* ⇒ ѓюѓюли.

Randomly applying the transliteration rules for the word "ghiocei", the following variants are obtained: {"гхиокеи", "гхиокеь", "гхиокей", "гхиочеи", "гхиочеь", "гхиочей", "гиокеи", "гиокеь", "гиокей", "гиочеи", "гиочеь", "гиочей", "ѓхиокеи", "ѓхиокеь", "ѓхиокей", "ѓхиочеи", "ѓхиочеь", "ѓхиочей"}.

In Figure 4 we show the transliteration scheme of the word "ghiocei". The scheme highlights the correct variant – "гиочей". The backtracking method allows eliminating wrong options step by step and

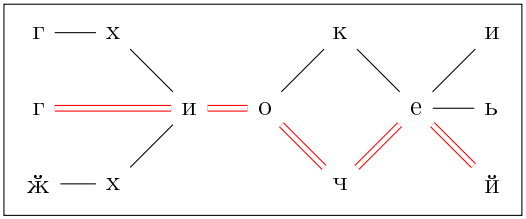


Figure 4. The transliteration scheme for "ghiocei"

selecting the correct one. This is done by changing the transliteration rule set, establishing some contextual dependencies, changing the order of the rules application and examining the MSD labels. In some situations the correct option can only be selected by the expert.

To fix the situations with multiple variants, we will use a list of options denoted by $[w_1][w_2]\dots[w_n]$, finally being selected only one. For example, for the words "ghiocei" and "preaiubiți" we get:

[гх][г][жх] • ио • [к][ч] • е • [и][ь][й] $\Longrightarrow$"гиочей",

пр • [еа][я]• [иу][ю] • биц • [и][ь][й] $\Longrightarrow$"пряюбиць".

We denoted by "•" the concatenation operation.

8 Classification of transliteration rules

8.1 General rules

The general transliteration rules are presented in Table 1.

Table 1. General rules of transliteration

latin	a	ă	â	b	c	d	e	f	g	h	i	î	j	k
cyrillic	а	э	ы	б	к	д	е	ф	г	х	и	ы	ж	к
latin	l	m	n	o	p	r	s	ș	t	ț	u	v	x	z
cyrillic	л	м	н	о	п	р	с	ш	т	ц	у	в	кс	з

To formalize (program) these substitutions, we will introduce the function *replace-all (w lat cyr)*, which will modify the word w substituting all occurrences *lat* with *cyr*. This is possible because the order

of application of these substitutions is not relevant. E.g, *replace-all* ("dividend" "d" "д") = "дividенд".

Depending on the filtering stage, it is possible to enter some new general transliteration rules, for example, *replace-all* (w "gh" "г"), *replace-all* (w "ch" "к").

Usually, these substitutions are the last filter in the process of cyrillization.

8.2 Rules for prefixes

Because the words are interpreted as strings, we have to use the notions of prefix and suffix defined to process strings, as opposed to the grammatical notions of suffix and prefix. Thus, by prefix (suffix) of the string w we will define any substring w_1 (w_2) for which $w = w_1 \bullet w_2$. Substrings w_1 and w_2 can also be empty, i.e. "". Often the transliteration rules for prefixes differ from general rules. Thus, the prefixes "ia" and "iu", with small exceptions, will be transliterated as "я" and "ю", as opposed to their appearance inside the word when in most cases they will be transliterated as options "[я][иа]" and "[ю][иу]". Another example: in the LexROM there are about 650 words that start with the prefix "crea". Only for 6 situations it will be transliterated by "кря". All other occurrences of the prefix will be transliterated by креа". These 6 situations can be easily highlighted and formalized. This observation suggests the need to introduce a special set of transliteration rules for prefixes.

We note these rules by *replace-prefix (w prefixlat prefixcyr)*. For example, *replace-prefix (w "creang" "крянг")*, *replace-prefix (w "crea" "креа")*. The order of application is important for this type of substitution, which is simple: prefixes that are prefixes of other prefixes will be transliterated last. Thus, we first will try the transliteration "creang" ⇒ "крянг", then the transliteration "crea" ⇒ креа". Prefix rules are defined separately for all words in the LexROM that begin with the same letter. Thus, for all letters there will be defined sets of rules for prefixes that will be applied first in the transliteration process.

8.3 Rules for suffixes

Analogously to the situation with the transliteration of prefixes, also there are defined transliteration rules for suffixes involving some specific conditions. For example, for the termination "ci" transliterations are possible: "ci" ⇒ "ч", "ci" ⇒ "чь", "ci" ⇒ "чи". To make the correct decision, MSD labels are checked. The rule "ci" ⇒ "ч" is applied, for example, for masculine nouns to the singular, nominative-accusative case (MSD = "Ncmsrn", "arici" ⇒ "арич", "cinci"⇒" чинч").

The "ci" ⇒ "чи" rule is applied to infinitive verbs and 3rd person verbs (MSD = "Vmis3s" and MSD = "Vmn", for example, "a munci" ⇒ "а мунчи"), and the rules "ci" ⇒ "чь", "ti" ⇒ "ть", "ți" ⇒ "ць", "și"⇒" шь", etc. – for nouns and adjectives in the plural dative-genitive case, and also for second-person present and past tense verbs (MSD ∈ {"Vmii2p", "Vmis2s", "Vmis2p", "Vmil2s", "Vmil2p", "Vmip2s", "Vmip2p", "Vmsp2s", "Vmsp2p", "Vmmp2p" }). Some examples are presented in Table 2.

Table 2. Transliteration of verb terminations

MSD	lat	cyr
Vmii2p	citeați	читяць
Vmis2s	citiși	читишь
Vmis2p	citirăți	читирэць
Vmil2s	citiseși	читисешь
Vmil2p	citirăți	читирэць

MSD	lat	cyr
Vmip2s	citești	читешть
Vmip2p	citiți	читиць
Vmsp2s	citești	читешть
Vmsp2p	citiți	читиць
Vmmp2p	citiți	читиць

Based on the above, unconditioned and conditioned rules are defined for the transliteration of suffixes. Respectively, the functions are defined *replace-suffix (w lat* cyr) and *replace-suffix-if (w label lat* cyr *msd*). The "label" argument of the *replace-suffix-if* function represents the label MSD of the processed word *w*, and the argument "msd" – a set of valid MSD labels for this rule. Unlike the rules for prefixes that are defined separately for each letter, the rules for suffixes are universal and can be applied to all words.

8.4 Context sensitive rules (for diphthongs and triphthongs)

As in the case of prefixes (suffixes), along with the usual grammatical notions diphthong/triphthong, we examine other combinations consisting of two, three or more letters. We mentioned above the behavior of the diphthongs "ia" and "iu" as prefixes, but also as occurrences within the word. Other examples are presented in Table 3.

Table 3. Transliteration of diphthongs/triphthongs

Diphthong/ triphthong	Transli- teration	Examples
"ioa"	[оа]	cioară⇒ "чоарэ"
	[ьоа]	chioară⇒ "кьоарэ"
	[иоа]	mioară⇒ "миоарэ"
"ii"	[ий]	fiicele⇒ "фийчеле"
	[ии]	viile⇒ "вииле"
"eie"	[ее]	creier⇒ "креер"
	[ей]	conveier⇒ "конвейер"

Diphthong/ triphthong	Transli- teration	Examples
"ea"	[а]	ceață⇒ "чацэ"
	[я]	rea⇒ "ря"
	[еа]	ocean⇒ "очеан"
"ch"	[к]	ochi⇒ "окь"
"gh"	[г]	ghid⇒ "гид"
"ge"	[же]	ger⇒ "жер"
"ci"	[чи]	circ⇒ "чирк"

Namely the transliteration of these constructions generates the most ambiguities. More information on this topic can be found in [12]. To make right decisions, sometimes contextual rules can be supplemented with morpho-syntactic information (MSD labels). The order of application of the rules is very important. Contextual dependencies always have priority over general rules.

9 Cyrillization algorithm

The cyrillization algorithm applies consecutively the transliteration rules, previously defined, to all the words in the LexROM. It is important to follow the order of application of the rules. Of course, optional combinations will be generated, which correspond to the ambiguities.

This means that later it is necessary to modify the transliteration rules or to request the intervention of the linguistic expert.

Below we present the formalized algorithm.

CYRILLIZATION ALGORITHM

0. ***Start***

1. The lexicon of the modern Romanian language LexROM and transliteration rules are given.
* We will build the Romanian Cyrillic lexicon LexCYR for the period 1967-1989 *\

2. Initial LexCYR $= \varnothing$, LexROM$_1$ = LexROM.

3. ***loop for all*** letters $\alpha \in$ AlphaROM ***do***

3.1. ***loop for all*** words $w \in$ LexROM$_1(\alpha)$ ***do***

3.1.1. Transliteration rules for prefixes are applied.

3.1.2. Transliteration rules for suffixes are applied.

3.1.3. Context-sensitive rules for transliteration are applied.

3.1.4. General rules for transliteration are applied. The obtained result is denoted by *wcyr*.

3.1.5. *wcyr* is included in LexCYR.

3.2. ***end loop***

4. ***end loop***

5. ***Stop***

10 Decyrillization

Decyirillization faces the same problems as cyrillization. General and contextual rules are also defined. The general rules are relatively simple, for example, а $\Rightarrow$ *a*, р $\Rightarrow$ *r*, ю $\Rightarrow$ *iu*, ь $\Rightarrow$ *i*. If only the general rules are applied to transliteration, we obtain, for example, пуюлуй $\Rightarrow$ *puiului*, бьет $\Rightarrow$ *biet*, боер $\Rightarrow$ *boer*, пепт $\Rightarrow$ *pept*. The last two transliterations are incorrect. Correct would be боер $\Rightarrow$ *boier*, пепт $\Rightarrow$ *piept*. In this case, as for cyrillization, contextual rules are required (for prefixes/suffixes, diphthongs/triphthongs). E.g, г$\bullet\beta \Rightarrow$ *gh*$\bullet\beta$, if $\beta\in${е,и,я,ю,ь} and г$\bullet\beta \Rightarrow$ *g*$\bullet\beta$, if $\beta\notin${е,и,я,ю,ь} (георгинэ $\Rightarrow$ *gheorghină*, гогоашэ $\Rightarrow$ *gogoașă*).

Rules for the letter я: я$\Rightarrow$ *ia* (usually at the beginning of the word), ия$\Rightarrow$ *ia* (usually at the end of the word). If it is difficult to make the

right decision to transliterate the letter я inside the word, then the algorithm will use the rule я $\Rightarrow [ia][ea]$, leaving the right decision to the expert. More information on this topic can be found in [13].

Another difficult problem is the transliteration of the letter ы, which can be replaced by either *î* or *â*. The algorithm follows exactly the recommendations of the Romanian Academy regarding this spelling.

DECYRILLIZATION ALGORITHM

0. ***Start***

1. The Romanian Cyrillic lexicon for the period 1967–1989 LexCYR and the decyrillization rules are given.
* We will build the lexicon of the modern Romanian language (noted by LexROM_2) applying the transliteration method *\

2. Initial $\text{LexROM}_2 = \varnothing$

3. ***loop for all*** letters $\beta \in$ AlphaCYR ***do***

3.1. ***loop for all*** words $w \in \text{LexCYR}(\beta)$ ***do***

3.1.1. Transliteration rules for prefixes are applied.

3.1.2. Transliteration rules for suffixes are applied.

3.1.3. Context-sensitive rules for transliteration are applied.

3.1.4. Transliteration rules for the letter кyr ы are applied.

3.1.5. General rules for transliteration are applied. The obtained result is denoted by *wrom*.

3.1.6. *wrom* is included in LexROM_2.

3.2. ***end loop***

4. ***end loop***

5. ***Stop***

11 Lexicon generation technology

As it was mentioned above, there is a total lack of electronic resources for the period 1967-1989, a complete exposition of the grammar used is missing, and many of interpretations of the transliterated words are ambiguous. Therefore, a major role in the process of generating the lexicon belongs to expert. The proposed technology aims to automate this process. Having the cyrillization and decyrillization algorithms

and the formalized sets of transliteration rules, the lexicon generation process can be realized as an backtracking algorithm. The process runs in several iterations, at each iteration the expert intervenes to modify the set of rules and, possibly, directly the built Cyrillic lexicon. This scheme is described in detail in Figure 5.

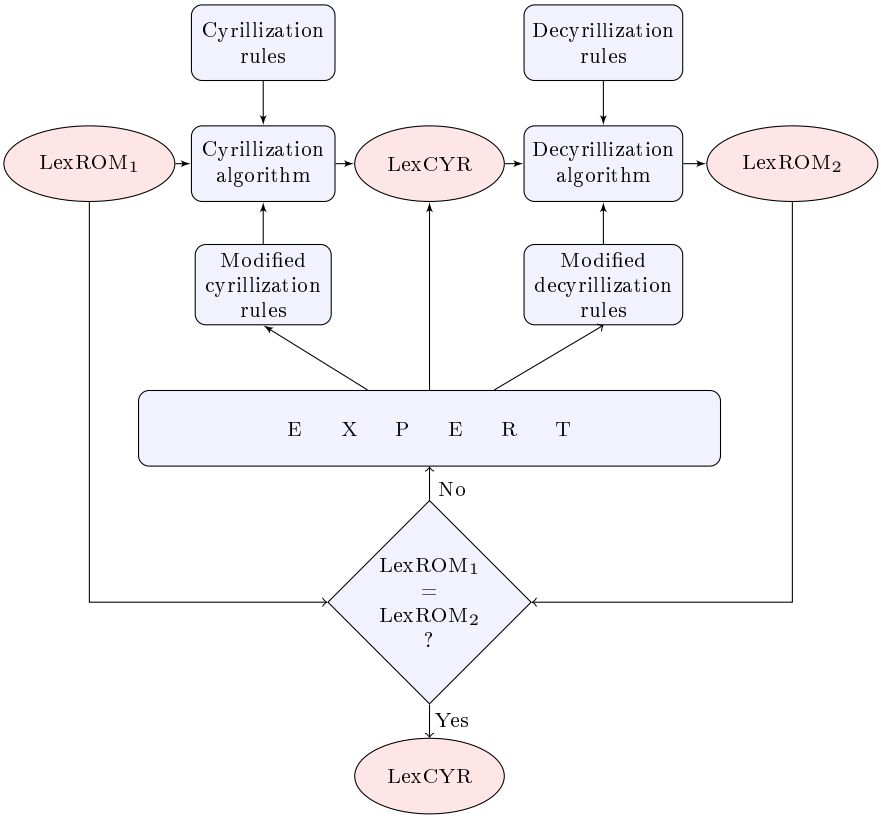


Figure 5. The scheme for generating the Romanian Cyrillic lexicon

12 Conclusion

The paper proposes a backtracking technology for the generation of the Romanian Cyrillic lexicon for the period 1967–1989 applying the

transliteration method. Starting from the lexicon of the modern Romanian language [6] the cyrillization and decyrillization algorithms are applied consecutively.

The intermediate results are made available to the experts, who can modify\extend the set of rules applied to transliteration, and to directly correct the built Cyrillic lexicon. The final lexicon is obtained as a result of performing several such iterations. The main problems to be solved by the experts are the ambiguities that appear as a result of cyrillization\decyrillization.

For all words in the LexROM(c) (171846 words), 6381 ambiguities were detected at the first iteration, which represents 3.7%. To overcome these ambiguities there were required two iterations. Of course, the degree of accuracy depends considerably on the qualification of the expert. The proposed technology allows the return to the previous intermediate variants, thus revising the lexicon.

In order to become aware of the role of the expert and that of contextual dependencies, a test was performed applying to the LexROM(c) only the general rules of transliteration (paragraph 8.1). As a result, 42.2% of the correct words are obtained.

References

[1] S. Cojocaru, E. Boian, C. Ciubotaru, A. Colesnicov, V. Demidova, and L. Malahov, "Regeneration of printed cultural heritage : challenges ang technolologies," in *The Third Conference of Matematical Society of the republic of Moldova*, (19-23 August, Chişinău), 2014, pp. 481–489.

[2] C. Ciubotaru, A. Colesnicov, and L. Malahov, "Vitalization of Moldavian Printings (1967–1989)", in *Proceedings of the 4th Conference of Mathematical Society of Moldova, CMSM4'2017"*, (June 28-July 2, 2017, Chisinau, Rep. of Moldova), pp. 491–494, ISBN 978-9975-71-915-5, Available: `http://cmsm4.math.md/Proceedings_CMSM4.pdf`.

[3] C. Ciubotaru, S. Cojocaru, A. Colesnicov, V. Demidova, and L. Malahov, "Regeneration of cultural heritage: problems related to

moldavian cyrillic alphabet", in *Proceedings of the 11th International Conference "Linguistics Resources and Tools for Processing the Romanian Language (ConsILR-2015)"*, (Alexandru Ioan Cuza University, Iași, Romania, 26-27 November 2015), pp.177–184, ISSN: 1843-911X, Available: `http://consilr.info.uaic.ro/2015/Consilr_2015.pdf`.

[4] *Dexonline. Online Dex. Romanian language dictionaries.* [Online]. Available: `https://dexonline.ro/definitie/translitera\%C8\%9Bie`.

[5] C. Ciubotaru, V. Demidova, and T. Bumbu, "Generation of the Romanian Cyrillic lexicon for the period 1967 – 1989," in *Proceedings of the Fifth Conference of Mathematical Society of Moldova IMCS-55*, (September 28 - October 1, Chisinau, Republic of Moldova), 2019, pp. 309–316.

[6] The UAIC Natural Language Processing Group, *WebPosRo/resources/*, Alexandru Ioan Cuza University, Faculty of Computer Science. [Online]. Available: `http://nlptools.info.uaic.ro/WebPosRo/resources/posDictRoDiacr.txt`.

[7] Tomaž Erjavec, ed., *MULTEXT-East Morphosyntactic Specifications, Version 3.0*, May 10th, 2004. [Online]. Available: `http://nl.ijs.si/ME/Vault/V3/msd/html`.

[8] *Reusable Resources for Romanian Language Technology*, Institute of Mathematics and Computer Sciences, Moldova. [Online]. Available: `http://www.math.md/elrr/res_main.php`.

[9] Guy L. Steele, *Common Lisp the Language*, 2nd ed., USA: Thinking Machines, Inc. Digital Press, 1990, 1029 p. ISBN:1-55558-041-6.

[10] *CLISP – an ANSI Common Lisp*, Slashdot Media. [Online]. Available: `http://sourceforge.net/projects/clisp/files/clisp/2.49/`.

[11] *Notepad++. Downloads.* [Online]. Available: `https://notepad-plus-plus.org/download/v7.7.1.html`.

[12] V. Demidova, "Particular Aspects of the Cyrillization Problem," in *The Third Conference of Matematical Society of the Republic of Moldova*, (Chişinău, 19-23 August), 2014, pp. 493–498.

[13] V. Demidova, "Peculiarities of decyrillization of the Romanian language," *Studia universitatis Moldaviae. Seria "Științe exacte și economice"*, no. 2(82), pp. 16–20, 2015. (in Romanian).

Constantin Ciubotaru

Vladimir Andrunachievici Institute of
Mathematics and Computer Science
Republic of Moldova
E–mail: chebotar@gmail.com

Part 3

Intelligent information system structures, databases, and knowledge bases for medical triage and diagnostic applications

An approach to structure information regarding patient diagnostics in the form of taxonomy in management of mass casualty disasters

Constantin Gaindric, Olga Popcova, Sergiu Puiu, Iulian Secrieru, Elena Gutuleac, Svetlana Cojocaru

Abstract

Processing of poorly structured data and knowledge remains very important, as processing methods greatly depend on the application domain. It is particularly difficult to provide activities with data, information and knowledge of good quality. This article presents main features of ill-structured problems, and describes the experience of structuring information regarding patient diagnostics in the form of taxonomy (based on vital and sonographic signs) in the process of management of mass casualty disasters. Using the described approach, during the working sessions of expert and "knowledge engineer" the task of acquisition and formalization of facts and decision rules was performed.

Keywords: ill-structured problems, knowledge acquisition, data structuration, taxonomy, mass casualty situation, sonographic diagnostics.

1. Introduction

The existence of society today, but even more so in the future, depends on information and communication technologies, which are integrated with the traditional ones, gradually replacing them.

Integrated technologies are increasingly applied in the real production environment, as well as in the artificial one, but also in services. Information technologies ensure fast and easy transfer of data in the form of text, images and voice, making services more efficient. Virtual reality

will make possible interactive virtual modelling and design of information systems at a more advanced level.

Predominantly unstructured nature of the data, on the basis of which the information systems operate, with those databases are populated, and which are traditionally manipulated, remains a permanent actual problem. Therefore, the processing of poorly structured data and knowledge remains important, especially since these processing methods at present, but also in the near future, greatly depend on the application domain. It is particularly difficult and important to provide activities with data, information and knowledge of good quality.

In order to improve the quality of data, information and knowledge in the daily activity and for the information systems, there is a need of new solutions and specific tools to be used.

Recently more and more attention is paid to the quality of data and information. Many researchers, including Pierce, Kahn, and Melkas [1], examine the relationship between data quality, information quality, and knowledge quality. It is stated that improving the data quality should lead to an improvement in the information quality, generated from this data. Therefore, it seems reasonable that improving information should, in its turn, improve the knowledge quality.

We are convinced that information systems will become the predominant factor in the progress of any field of activity, including in the help of doctors, especially to provide a correct, fast and efficient diagnostics.

2. Features of ill-structured problems

Ill-structured problems are the ones that everyone commonly faces in his/her everyday life. These include important social, political, economic and scientific issues.

The solution for ill-structured problems, usually, requires the following activities: a) definition, description, problem formalization; b) generating possible solutions; c) evaluation of alternative solutions, taking into account the end-users' preferences; d) implementation of the most viable solution; e) monitoring of the implementation.

Knowing the domain and its good description are the main factors in solving ill-structured problems. In addition, professional skills and

knowledge, involved in generating solutions in the decision-making process, should be identified. All solutions should also have the justification component in order to be evaluated later on. Other two important components of solving ill-structured problems are: taking into account the decision maker's view and selecting the solution based on a personalized end-user approach.

Usually, solutions for ill-structured problems are rarely correct or incorrect, but they should fall within a range of acceptances. As a result, in order to be judged there are needed the stages of testing, implementation and evaluation based on the arguments.

The stages of solving ill-structured problems are very similar with the five stages of the development of knowledge-based systems: 1. Problem identification; 2. Knowledge acquisition; 3. Knowledge structurization; 4. Knowledge formalization; 5. Prototype development, testing and implementation. This is due to the fact that both processes are based on human reasoning logic.

As in the case of the development of knowledge-based systems, the main stage in solving ill-structured problems is considered the phase of knowledge acquisition, structurization and formalization. The aim is to obtain an informal description of the knowledge regarding the studied domain in the form of a graph, table, diagram or formatted text.

Taxonomy is the most commonly used form of representation of the structured description of the problem area.

The following section describes the experience of structuring information regarding patient diagnostics in the form of taxonomy (based on vital and sonographic signs) in the process of management of mass casualty disasters, an eloquent example of a domain with poorly structured and heterogeneous data and knowledge.

3. Knowledge structurization and formalization in the form of taxonomy during expert-"knowledge engineer" interaction

As quality of the acquired knowledge is the determining factor for the successful solution of any ill-structured problem, the decision was made to use the traditional method of knowledge acquisition – with participation of the "knowledge engineer" [2]. The main objective of this method is that during the sessions of professional knowledge acquisition, the competence

and expert knowledge should be transferred to the "knowledge engineer", in order to obtain the most complete possible representation of the problem area.

Later, the information obtained from medical experts was structured, formalized and introduced by the "knowledge engineer" into the knowledge base (a pyramid of meta-concepts and a set of rules created on their basis), using the ExpShell tool [3].

The main characteristics/notions (vital and sonographic signs) used in the description of the domain of patient diagnostics in the process of management of mass casualty disasters have formed the main nodes. The other notions have formed the nodes of higher levels, being connected to the main ones through hierarchical links – in this way forming a tree structure of "attribute" and "value" nodes.

The obtained hierarchical structure is nothing but a taxonomy of the problem area.

Common work of the "knowledge engineer" and experts has shown that in medical examination domain the reasoning with meta-concepts (facts) and knowledge representation as a pyramid/taxonomy completely corresponds to the experts' mentality and thinking.

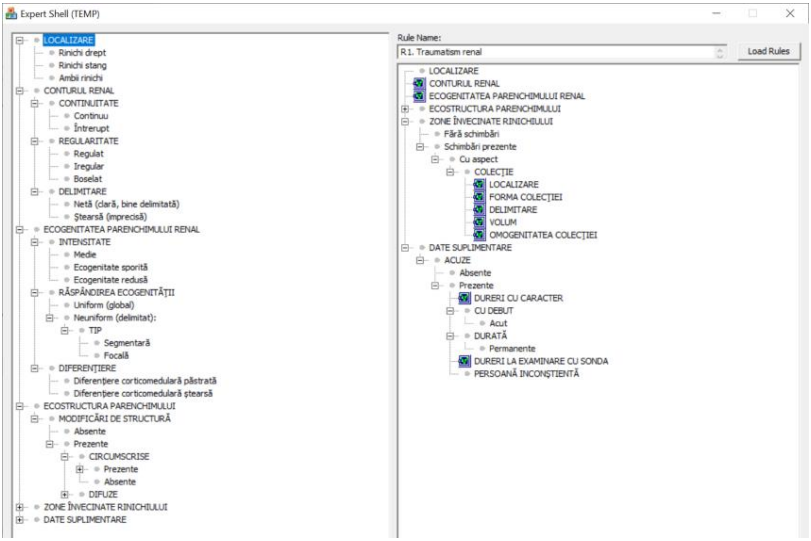


Figure 1. Domain formalization – kidneys (prototype).

The domain of "mass casualty disasters" was studied as a domain with poorly structured and heterogeneous data and knowledge. In particular, information about the injuries of abdominal organs (liver, pancreas, kidneys, and spleen) was acquired. The acquired information was structured in the form of facts (see Figure 1) – preparatory action for formalization and creation of decision rules. Based on these facts the decision rules were created. Later, the created decision rules were validated.

Information about fluid volume and thoracic air volume, based on features captured by sonographic scanners, were acquired. The acquired information was structured in the form of facts. The structure of facts was revised, in according to the Extended Focused Assessment with Sonography in Trauma (EFAST) (see Figure 2), an extended version of FAST protocol, used for sonographic examination in case of mass casualty situations [4-5].

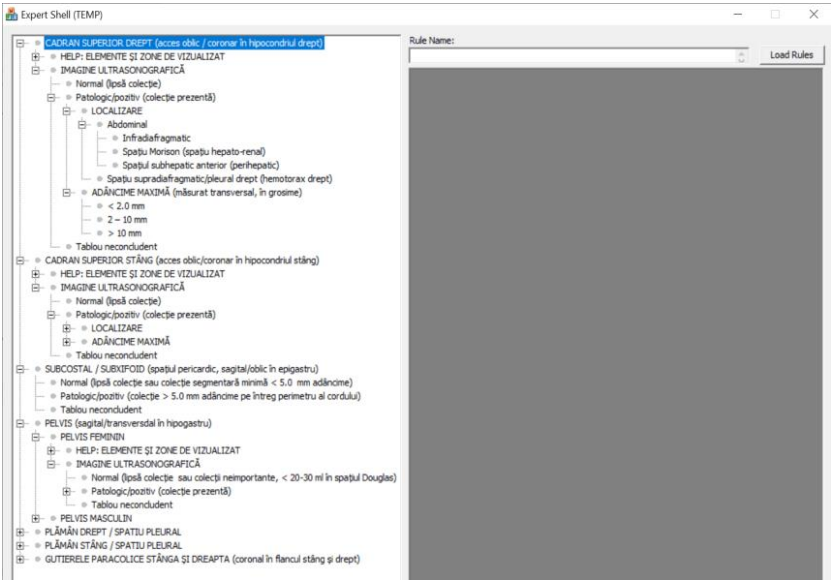


Figure 2. EFAST protocol (prototype).

Information about casualty state based on the vital signs (state of consciousness, pulse, respiratory rate, blood pressure) was acquired. The

acquired information was structured in the form of facts (see Figure 3). Based on these facts the decision rules for 4 classes of hemorrhagic shock (Class 1-4) were created. Later, the created decision rules were validated.

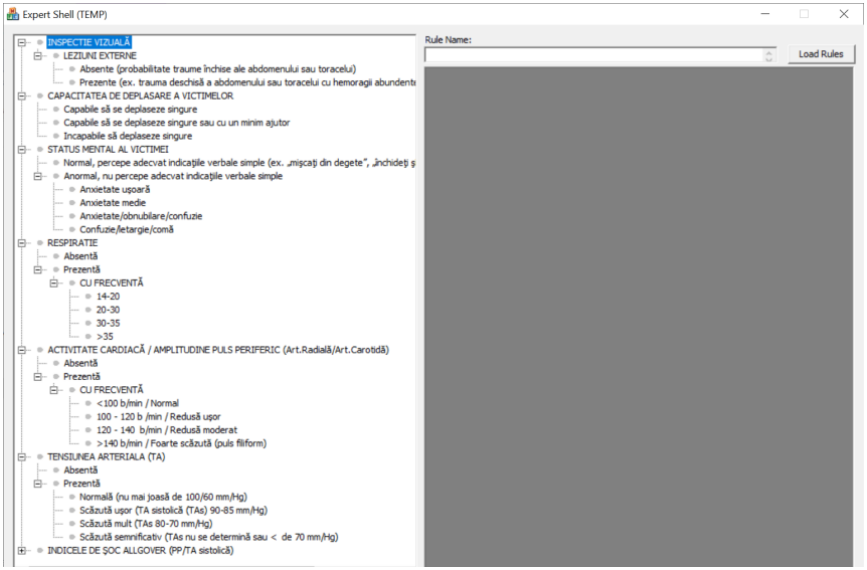


Figure 3. Casualty state based on the vital signs (prototype).

As a result, a prototype of the formalized domain was created.

Table 1. Knowledge base results

	Root nodes	Main nodes	Depth levels
Liver	5	129	10
Pancreas	6	149	10
Kidneys	6	115	11
Spleen	5	114	11
Shock	7	30	4
EFAST	7	63	6

5. Conclusion

Within the described approach, information solutions were proposed, with an increased level of intelligence, to assist the decision makers of ill-structured problems at the stage of problem definition, description and

formalization. These solutions were based on methods and algorithms in the fields of decision support systems, knowledge-based systems, advanced methods of new professional knowledge acquisition and identification, logical inferences oriented on ill-structured problems, and elements of artificial intelligence.

The proposed solution takes into account the fragmentary and heterogeneous structure of information, data and knowledge in the problem area – patients diagnostics in the process of management of mass casualty disasters. This involves studying how to integrate different data sources by using the taxonomies/ontologies associated with these data sources in order to define standardized structures, providing interoperability and consistency of stored data.

The implementation of the proposed approach allows one to follow the reasoning of the decision maker(s) from the initial stage until the solution of the concrete case/precedent. In the case of the solution confirmation and its argumentation, it can be disseminated as "good practice" or "malpractice" – a very important issue in poorly structured domains.

Acknowledgments. The research for this paper have been supported by the project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" in the framework of the State Program and G5700 NATO Science for Peace and Security Programme.

References

[1] E. Pierce, B. Kahn, H. Melkas. *A comparison of quality issues for data, information, and knowledge.* In: M. Khosrow-Pour (Ed.), Emerging Trends and Challenges in Information Technology Management: Proceedings of the 2006 Information Resources Management Association Conference, 17th IRMA International Conference, Washington, DC, USA, May 2006, pp. 21-24.

[2] L. Burtseva, S. Cojocaru, C. Gaindric, E. Jantuan, O. Popcova, I. Secrieru, D. Sologub. *SONARES – A decision support system in ultrasound investigations.* Computer Science Journal of Moldova, vol. 15, nr. 2 (44) (2007), pp. 153-177.

[3] Iu. Secrieru, D. Sologub. *Expert shell aimed at creation of the knowledge base for ultrasonic research intelligent system.* Revista de inventica, nr. 48, vol. IX (XV-2005), pp. 7-12.

[4] American Institute of Ultrasound in Medicine, American College of Emergency Physicians. *AIUM practice guideline for the performance of the focused assessment with sonography for trauma (FAST) examination*. J Ultrasound Med, Nov. 33 (11) (2014), pp. 2047-2056.

[5] A.W. Kirkpatrick, M. Sirois, K.B. Laupland, D. Liu, K. Rowan, C.G. Ball, et al. *Hand-held thoracic sonography for detecting post-traumatic pneumothoraces: the Extended Focused Assessment with Sonography for Trauma (EFAST)*. J Trauma, 57(2) (2004), pp. 288-295.

Constantin Gaindric[1], Olga Popcova[2], Sergiu Puiu[3], Iulian Secrieru[4], Elena Gutuleac[5], Svetlana Cojocaru[6]

[1]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: constantin.gaindric@math.md

[2]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: oleapopcova@yahoo.com

[3]Affiliation/Institution: Medical Center "Ana Maria", State University of Medicine and Pharmacy "Nicolae Testemitanu"
E-mail: puiusv@yahoo.com

[4]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: iulian.secrieru@math.md

[5]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: elena.gutuleac@math.md

[6]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: svetlana.cojocaru@math.md

Introducing an AI-based Response Framework for Mass Casualty Management

Marian Sorin Nistor, Van Loi Cao, Truong Son Pham,
Stefan Pickl, Constantin Gaindric, Svetlana Cojocaru

Abstract

Advances made in Artificial Intelligence over the last couple of years have revealed certain limitations to traditional Mass Casualty Management (MCM).

This paper introduces a MCM response framework using state-of-the-art OR-based models of existing AI solutions aimed to optimize each stage (extrication, triage, and transportation) of the response phase during a Mass Casualty Incident (MCI).

Keywords: Mass Causality Management, AI-based solutions, extrication, triage, and transportation.

1. Introduction

An MCI refers to a multiple-casualty situation in which Emergency Medical Services (EMS) resources, such as personnel and equipment, are overloaded. This occurs when the number of casualties is often much higher than the available resources in a particular area [1]. MCI events often result from transportation accidents, terrorism, fires, or natural disasters.

MCM is a widely used model for managing victims in MCIs. It often employs a multi-sectional approach for managing the strong connection between triage, field stabilization, and evacuation [2]. There are two key components to this approach: (1) a command post that coordinates the incorporated links between the field and health care facilities, (2) levels of special knowledge attributed to responders (e.g., policeman, firefighter, search and rescue, pre-hospital team).

2. Key Components in MCM

An optimal MCM should consider the four phases of the disaster cycle: mitigation, planning, response, and recovery [1].

In the first phase, some of the devastating effects of disasters can be reduced by acting before the actual event happens. In the second phase, realistic disaster planning and practicing are finally revised. The response consists of a series of necessary procedures from notification of an MCI, searching, rescuing, and sending casualties to hospitals. Finally, the recovery phase rebuilds and reconstructs the infrastructure while taking some actions to reduce future disasters.

This paper focuses on the response phase, more specifically on extrication, triage, and transportation. It is a sequential process from collecting and classifying victims to sending them to appropriate hospitals. Extrication is the process that prioritizes patients based on the severity of their conditions for further actions, such as immediate movement or treatment. Triage aims to provide the most efficient aid to as many casualties as possible and prioritizes treatment and transportation of victims. This requires a dynamic balance between needs (types and number of injuries) and resources (infrastructure, equipment, and competent personnel ability).

During this stage, patients are triaged and transported based on different variables, such as the number of victims, the type of incident, the available resources, and the existing infrastructure capability. Thus, triaging errors may lead to worse outcomes in later processes and an increasing number of fatalities.

The patients are typically tagged as red (critical cases - major life-threatening injuries), yellow (urgent cases, non-threatening injuries), green (non-urgent cases), and black (unlikely survival).

3. AI-based Decision Support Systems for MCM

Recently, AI is applied to build decision-support tools used in the various phases of MCM models [4-7]. The AI-based decision support tools aim to assist in resource management for disaster response with a broad range of objectives and decision variables.

One usage is, e.g., in the search and rescue phase, where Mishra et al. [8] proposed a state-of-the-art detection method based on computer vision and developed a large dataset for searching and rescuing in natural disasters using drone surveillance; Perry et al. [9] introduced a triage method based on computer vision to provide real-time casualties information at the disaster scene for the MCI commander and the EMS dispatch.

Moreover, effective decisions regarding the evacuation of mass casualty patients to hospitals should involve the assessment of the facility capacity, such as the availability of beds and the ability to deal with the patient overload. Information regarding real-time hospital bed capacity is key to controlling the flow of patients from an MCI. In this case, it allows for evidence-based decision making regarding the evacuation of patients.

Various decision support models are proposed for resource management in disaster response [3]. Several factors should be considered for integration into the triage process, such as available resources and the scale of the disaster.

There are three categories of AI-based solutions for MCM models [3]: traditional optimization-based decision support, AI-based optimization, AI-based transportation congestion detection. However, these solutions are often proposed to address issues of different components in the MCM response framework. There is no MCM response framework supported by AI-based solutions for almost all its components, including the MCI area, triage stage, and transportation stage, to the best of the authors' knowledge. Therefore, this paper introduces an MCM response framework with extended support from recent AI-based solutions.

4. Proposed Response Framework for MCM

The recent developments of AI-based support tools for MCM can be employed in one MCM model. The proposed framework aims to inherit the advantages of different AI-based solutions that have been proposed for MCM separately. The overview of the proposed framework is described in Figure 1. In the following, the AI-based solutions for the components of the response framework are discussed.

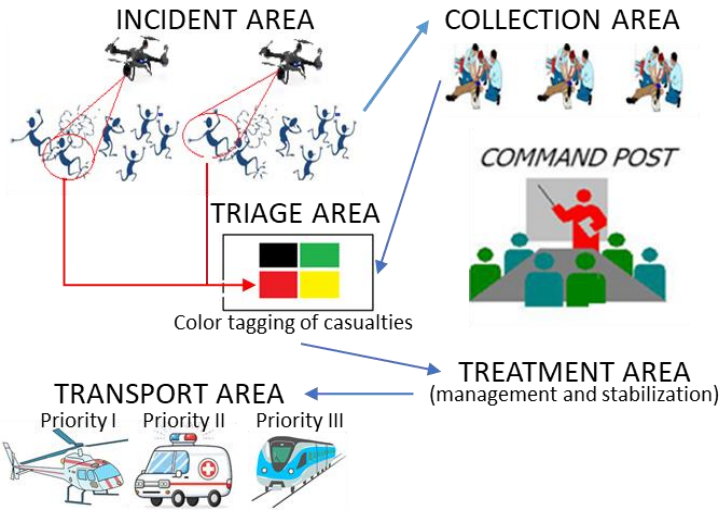


Figure 1. Response framework for an optimal MCM

(1) Search and rescue at the incident site: the task can be performed by applying the detection method introduced by Mishra et al. [8] or the triage method of Perry et al. [9]. In these methods, the AI-based detection models can analyze the video captured by drone surveillances to search for casualties. However, the AI-based triage method et al. [9] can be further used in the triage area.

(2) Search and rescue in natural disasters (i.e., floods): in this scenario, the victims are distributed in a large region. The search and rescue teams collect the casualties based on their requests. An optimization-based model proposed by Yan et al. [6] can be applied. In this model, the Support Vector Machine (SVM) and Reinforcement Learning (RL) work together to estimate the density of the regions of victim requests. Every time a victim is aided, the RL algorithm updates the density of patient requests.

(3) Triage process: the category information of each victim produced by the AI-based triage method of Perry et al. [9] for search and rescue automatically tags them into appropriate triage areas. Casualties with life-threatening injuries could ignore the collection area and be sent directly to triage for the quick delivery to hospitals. Here, expert knowledge is needed for final confirmation.

(4) Transportation: many AI-based models are proposed for real-time traffic congestion detection, as reviewed in [3]. These models typically use deep learning methods, such as Convolutional Neural Networks (CNNs), to identify vehicles and estimate their density on a given road segment. Among the reviewed models, the one proposed by Du et al. [10] is very promising. Besides traffic information, the model incorporates several factors, such as weather conditions, flying attitude of drones or unmanned aerial vehicles (UAVs), and vehicle category, to improve the detection accuracy.

5. Conclusion

AI is employed to address many phases in MCM, predominantly components in the MCM response phase. AI-based solutions, however, are often applied to address problems separately in MCM models. This paper introduces an MCM response framework in which most of its components are supported by robust AI-based models proposed in recent years.

Acknowledgments: This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700. Proofreading of Jacob Ehrlich is gratefully acknowledged.

References

[1] Y. Kluger, F. Coccolini, F. Catena, and L. Ansaloni. *WSES Handbook of Mass Casualties Incidents Management*. Hot Topics in Acute Care Surgery and Trauma Ser. Springer, Cham, 2019.

[2] Pan American Health Organization: *Establishing a Mass Casualty Management System*. PAHO, 2001.

[3] Marian Sorin Nistor, Truong Son Pham, Stefan Pickl, Constantin Gaindric, and Svetlana Cojocaru: *A concise review of AI-based solutions for mass casualty management.* In Proceedings of the 1st International Workshop on Computational & Information Technologies for Risk-Informed Systems (CITRisk-2020), October 16-18, 2020, Kherson, Ukraine. Accepted.

[4] B. Kamali, D. Bish, and R. Glick. *Optimal service order for mass-casualty incident response*. EJOR 261, pp. 355–367, 2017.

[5] D.T. Wilson, G.I. Hawe, G. Coates, R.S. Crouch. *A multi-objective combinatorial model of casualty processing in major incident response*. EJOR, 2013.

[6] L. Yan, S. Mahmud, H. Shen, Y. Foutz, and J. Anton. *MobiRescue: Reinforcement Learning based Rescue Team Dispatching in a Flooding Disaster*. In Proc. ICDCS. IEEE, Singapore, July 8-10, 2020.

[7] Q. Meng, H. Song, Y. Zhang, X. Zhang, G. Li, and Y. Yang. *Video-Based Vehicle Counting for Expressway: A Novel Approach Based on Vehicle Detection and Correlation-Matched Tracking Using Image Data from PTZ Cameras.* Mathematical Problems in Engineering, 2020.

[8] B. Mishra, D. Garg, P. Narang, and V. Mishra. *Drone-surveillance for search and rescue in natural disaster*. Computer Communications, 2020.

[9] O. Perry, A. Bar-Hillel, and Y. Bitan. *Using Deep Learning to Provide Real Time Information During Mass-Casualty Incident*. In Proc. IPRED, 2020.

[10] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. *The unmanned aerial vehicle benchmark: Object detection and tracking.* In Proc. ECCV, pp. 370–386, 2018.

Marian Sorin Nistor[1,a], Van Loi Cao[2], Truong Son Pham[1,b], Stefan Pickl[1,c], Constantin Gaindric[3,a], Svetlana Cojocaru[3,b]

[1]Department of Computer Science, Universität der Bundeswehr München, Germany

[a]E-mail: sorin.nistor@unibw.de

[b]E-mail: son.pham@unibw.de

[c]E-mail: stefan.pickl@unibw.de

[2]Department of Computer Science, Le Quy Don Technical University, Hanoi, Vietnam

E-mail: loi.cao@lqdtu.edu.vn

[3]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chisinau, Republic of Moldova

[a]E-mail: constantin.gaindric@math.md

[b]E-mail: svetlana.cojocaru@math.md

Regional intelligent data warehouse for DLD cases

Iulian Secrieru, Elena Gutuleac, Olga Popcova

Abstract

Diffuse liver diseases (DLD) is a continuing concern of internal medicine and a subject of numerous research and publications. Correct and early assessment of liver diseases combined with appropriate management of pathologies can certainly increase the patients' quality of life and its duration. The main aim of the research is to formalize and integrate data and scientific knowledge from the fields of diagnostics and treatment of DLD, which at the moment are unstructured, fragmented and heterogeneous into a unique informational space. Data and knowledge digital warehouse is used in order to allow the interoperability of the stored data contents and knowledge, for healthcare shareholders from EaP region (Moldova, Armenia, Azerbaijan, Belarus, Georgia and Ukraine).

Keywords: medical informatics, diffuse liver diseases, data digital warehouse.

1 Introduction

Chronic diffuse liver diseases play an important role in morbidity and mortality of the population of many economically developed countries, but also in developing and transition countries. Liver disease accounts for approximately 2 million deaths per year worldwide, one million due to complications of cirrhosis and one million due to viral hepatitis and hepatocellular carcinoma. Hepatic cirrhosis is currently the 11th most common cause of death globally and liver cancer is the 16th leading

cause of death; combined, they account for 3.5% of all deaths worldwide. Cirrhosis is within the top 20 causes of disability-adjusted life years and years of life lost, accounting for 1.6% and 2.1% of worldwide burden [1]. The burden of liver disease in Europe continues to grow [2]. We have witnessed the wide spread of DLD in the whole EaP region, which predominantly affect people of working age, having a significant negative impact on social and economic development of the countries [3].

Diagnostics of DLD requires special knowledge and use of laboratory and instrumental diagnostic methods, sometimes unique or rarely used in daily diagnostics. This fact causes impediments not only to novices, general practitioners, but also to experienced doctors. At the same time, international groups of experts have developed criteria, scales and diagnostic scores, based on certain parameters. Although these tools (criteria, scales and scores) are quite accurate, they are inconvenient in their use in everyday clinical practice.

There is a clear need to create and provide clinicians with information tools for collecting, storing and processing medical test data and instrumental investigations of patients predisposed or suspected of having a DLD.

2 Methods and Results

At the stage of elaboration of a single protocol for description of DLD the methods and practices from the domain of artificial intelligence (advanced technologies for acquisition of professional medical knowledge, structurization algorithms and creation of medical taxonomies, representation methods in the form of medical ontologies) have been used [4].

At the stage of information, data and expert knowledge collection and storage practices from the domain of relational databases have been used. For structurization and primary data processing classical methods of medical statistics have been used.

At the stage of development of tools for quantifying and assessing

DLD advanced algorithms from the domain of artificial intelligence, such as segmentation and clusterization, have been used [5].

Development of the user interface is based on existing practices, principles and approaches used in modern medical information systems. As a result, we created tools for quantifying and assessing DLD, which could be integrated in regional intelligent data warehouse.

The tools allow:

- to aggregate a large number of DLD cases in a standardized manner;
- to define criteria based on non-invasive measurements and laboratory tests for quantifying and assessing DLD;
- to establish thresholds and endpoints for onset and all stages of DLD progress.

As a regional data warehouse, we intend:

- to harmonize efforts of data stakeholders for in-depth DLD phenotyping;
- to promote clinically impactful new knowledge discovery and its translation into clinical practice.

3 Conclusion

The main criteria based on non-invasive measurements and laboratory tests for quantifying and assessing DLD have been identified. Based on these criteria taxonomy for DLD cases formalization in the frame of data warehouse have been created. Algorithms and informational tools for establishing the thresholds and endpoints for onset and all main stages of DLD progress have been developed. The regional intelligent data warehouse, as a computer-based shared platform, will intensify, simplify and facilitate knowledge exchange between healthcare shareholders from EaP region, that otherwise would be difficult or even impossible.

Acknowledgments. State Program Project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data"; G5700 NATO Science for Peace and Security Programme; EYR@EaP 2019 programme have supported the research for this paper.

References

[1] Asrani S.K., Devarbhavi H., Eaton J., et al. *Burden of liver diseases in the world.* Journal of hepatology, 2019.

[2] Pimpin L., Cortez-Pinto H., Negro F., Corbould E., Lazarus J.V., Webber L., Sheron N., Committee EHS. *Burden of liver disease in Europe: Epidemiology and analysis of risk factors to identify prevention policies.* Journal of hepatology. 2018; 69(3) pp. 718—735.

[3] *World Health Organization Regional Office for Europe.* European Detailed Mortality Database. 23 August 2016.

[4] Secrieru Iu. *Structured knowledge management techniques for the development of interactive and adaptive decision support system.* Computer Science Journal of Moldova, nr. 1 (49), vol. 17, 2009, Kishinev, pp. 58–73.

[5] Secrieru Iu., Popcova O., and Gutuleac E. *Quantification and Assessment of Diffuse Liver Diseases Using Deep Data Analysis.* Proceedings of the Conference on Mathematical Foundations of Informatics, Chisinau, Moldova, November 9-11, 2017, pp. 166–167.

Iulian Secrieru[1], Elena Gutuleac[2], Olga Popcova[3]

[123]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mails: [1]iulian.secrieru@math.md, [2]elena.gutuleac@math.md, [3]oleapopcova@yahoo.com

Advanced pre-hospital triage based on vital signs in mass casualty situations

Constantin Gaindric, Sergiu Șandru, Sergiu Puiu, Olga Popcova, Iulian Secrieru, Elena Guțuleac

Abstract

In case of disasters, the pre-hospital triage is a very important stage, crucial for providing the necessary medical assistance and facilitating casualty evacuation from the disaster site to the nearest specialized hospitals. When mass casualty situations take place, casualty prioritizing for every triage category (Red, Yellow, Green) could help to improve the distribution of the available resources (healthcare personnel, ambulances) and, finally, save more lives. This article describes how a computer-aided tool for pre-hospital triage based on vital signs can be used in practice on the disaster site, as well as its possible applications in training purposes.

Keywords: medical informatics, pre-hospital triage, mass casualty situations, knowledge acquisition, computer-aided tool.

1. Introduction

The efficiency of medical first aid, especially in mass casualty situations, is extremely dependent upon the time of provided treatment. In the disaster area, usually, the number of casualties could be extremely high, many of them require urgent medical assistance simultaneously, but the available capabilities and resources (such as healthcare personnel/aides and ambulances) are limited.

Therefore *pre-hospital triage* is one of the most important elements in crisis management response in large-scale disasters. Pre-hospital triage helps to classify rapidly casualties in various homogeneous categories, taking into account the severity of injuries, in order to provide efficient

medical assistance and evacuation from the impact zone to the nearest specialized hospitals.

To support the pre-hospital triage process, as a part of the decision support framework for the management of complex mass casualty situations [1], we are developing a computer-aided tool for pre-hospital triage. This tool is aimed to gather primary medical data of casualties, to assess the triage category via rule-based decision support, and to give the possibility of *setting-up a priority within every triage category*.

2. Methodology

Commonly the following 3 basic categories are used for casualties assessment in the triage methods and algorithms:

- Red (Absolute emergency) – Life-threatening casualties with serious and very serious injuries, illnesses, intoxication, or contamination, who require immediate stabilization measures, as well as priority evacuation in assisted medical transport conditions.
- Yellow (Relative emergency) – Casualties with serious or moderate injuries, illnesses, intoxication, or contamination, with retained vital functions, but with the risk of developing life-threatening complications immediately ahead. They require urgent medical assistance, but not an immediate one.
- Green (Minimal emergency) – Casualties with minor injuries, illnesses, intoxication or contamination, no life-threatening, which can be treated later, usually in outpatient conditions. They can be evacuated in non-specialized transport or independently.

Some triage approaches consider additional categories such as Orange [2] or Gray [3] in order to use more effectively the medical personnel. This becomes especially important under the resources scarcity or when the road infrastructure also was affected by disaster, making casualties transportation problematic.

The need for casualty prioritizing in every triage category (Red, Yellow, Green) in pre-hospital conditions seems to be an advanced step, helping to improve the distribution of the available resources and, finally, save more lives.

3. Designing a computer-aided tool for pre-hospital triage

We have studied and analyzed different clinical and emergency guides and protocols, including the national ones – for Moldova [4]-[5]. As a result, we have selected the following basic attributes (casualty characteristics) which determine the decisions for triage based on vital signs, given in Table 1, and allow quick categorization of casualties:

Table 1. Basic attributes and values in triage based on vital signs

	Red (I)	Red (II)	Yellow	Green
Glasgow Coma Scale	3-8	9-13	14	15
Airways Permeability	Obstruction / Stridor	Difficult breathing	Normal breathing	Normal breathing
Pulse	>120 or <40	111-120 or 41-45	81-110 or 46-59	60-80
Systolic Blood Pressure	<80	80-89	90-100	>100
Respiratory Rate	>35 or <13	29-35	19-28	14-18
Oxygen saturation	<= 85	86-90	91-95	96-100
Individual mobility	Unable	Unable	With help	Walking

The Glasgow Coma Scale is used to objectively evaluate a person's level of consciousness after an injury. Its assessment is based on three aspects of responsiveness: eye-opening, motor, and verbal responses.

All these attributes and values allow us to create decision rules to distinguish priority I and priority II in the Red triage category, supporting the *decision-making*. In addition, our computer-aided tool will provide the possibility to end-user (as an option) to set up a priority within every triage category, helping to follow the casualty more accurate status, avoiding under-triage and over-triage, and suggesting life-saving interventions for casualty, needed in every specific case (with some prioritizing in the chain of emergency care).

4. Conclusions and future work

The standard clinical protocol in emergency cases, designed mainly for a single or limited number of persons, is not suitable for direct use on-site in case of large-scale disasters. Also limitation of the number of triage groups only to three (Red, Yellow, and Green) can lead to some problems in case of mass casualty situations, having a lot of persons in Red and Yellow categories in a short period of time.

Therefore there is a need to advance the algorithm and reasoning for casualty prioritizing for every triage category (Red, Yellow, Green), based not only on theoretical knowledge but on practical experience.

The described approach for advanced pre-hospital triage based on vital signs in mass casualty situations represents the basis for the elaboration of different applications having multiple purposes:

- Being developed as a computer-aided tool on mobile devices, it can be used by the rescue teams members in practice on disaster sites.
- Being implemented as a web application, it can be used for both teaching and training of paramedics, and for their evaluation (determining the level of knowledge and practical skills).

Also, this application can be used to interact with experts for the acquisition of new knowledge (by analyzing non-ordinary cases and collecting data on possible different opinions for specific cases) or for validation of the detected cut-off thresholds, which can be used to stratify casualties.

If it is possible, prioritizing in the Yellow category is extremely desirable, of course, in case the needed resources are available.

The validation process will include obtaining explanations and interpretations of all conclusions, including intermediate ones, obtained in the process of using the proposed pre-hospital triage.

Acknowledgments. The Moldovan State Program project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data” and the project G5700 “An Adaptive Decision Support Framework for the Management of Mass Casualty via an Artificial Intelligence Based Multilayered Approach integrating an Intelligent Reachback Information System” in the

framework of the NATO Science for Peace and Security Programme have supported part of the research for this paper.

References

[1] C. Gaindric, S. Cojocaru, S. Pickl, S. Nistor, Iu. Secrieru, O. Popcova, D. Bein, and D. Cimpoesu. *A Concept for a Decision Support Framework for the Management of Complex Mass Casualty Situations at Distribution Points.* In: Proceedings of the Conference on Mathematical Foundations of Informatics MFOI'2018, July 2-6, 2018, Chisinau, Republic of Moldova, pp. 90-102.

[2] C. Barfod, M.M.P. Lauritzen, J.K. Danker, et al. *Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department – a prospective cohort study.* Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, vol. 20 (2012), article number: 28. DOI 10.1186/1757-7241-20-28.

[3] K. Sakanushi, T. Hieda, T. Shiraishi, et al. *Electronic triage system for continuously monitoring casualties at disaster scenes.* Journal of Ambient Intelligence and Humanized Computing, 4 (2013), pp. 547-558. DOI 10.1007/s12652-012-0130-2.

[4] Gh. Ciobanu, M. Pîsla, F. Gornea, et al. *Ghid naţional privind tirajul medical în incidente soldate cu victime multiple şi dezastre*. Centrul Nat. Şt.-Practic Medicină de Urgenţă, Centrul Republican Medicină Calamităţilor. – Chișinău, 2010, 36 p.

[5] M. Ciocanu, Gh. Ciobanu, V. Cojocaru, A. Oglinda, L. Chiosea, N. Buzatu, and I. Gurov. *Protocol clinic standardizat. Triajul în Unităţile Primiri Urgente*. Chișinău, 2017, 23 p.

Constantin Gaindric[1], Sergiu Șandru[2], Sergiu Puiu[3], Olga Popcova[4], Iulian Secrieru[5], Elena Guțuleac[6]

[1456]Affiliation: Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chisinau
E-mails: constantin.gaindric@math.md, oleapopcova@yahoo.com, iulian.secrieru@math.md, elena.gutuleac@math.md

[2]Affiliation: Emergency Medicine Institute, Chisinau
E-mail: serghei.shandru@gmail.com

[3]Affiliation: Medical Center "AnaMaria-Med", Chisinau
E-mail: puiusv@yahoo.com

Tools for Triaging in Mass Casualty Incidents

Olesea Caftanatov, Tudor Bumbu

Abstract

In this paper, we propose two tools for triaging process in mass casualty incidents. One tool is a mobile application that records, analyzes, and sets transportation order. The second tool is a web platform for management and storage purpose.

Keywords: mass casualty incidents, tools, triage processes, management platform.

1 Introduction

During any natural or man-made incident, which results in extraordinary levels of mass casualties, local and state medical personnel are often overwhelmed by the sheer magnitude of the situation. It was found that the biggest challenges to providing care are resources and communication restraints. Thus, the key to use the resources more efficiently is to keep patients moving toward definitive treatment through accurate triage of life threats.

The term triage originates from the French word trier, meaning *to sort*. In the medical context, this *"sorting"* is a method of selection and classification of patients, by priority, for initial treatment and subsequent transportation to a facility, where more definitive care is available. The classical triaging is made on paper forms, but completing them is a time consuming procedure. Additionally, classical triage has communication restraints. Due to modern technologies, we can digitize the triaging process, and this is our main purpose of this research.

2 Related Works

In the last decade, the development of tools for Mass Casualty Incidents (MCI) obtained great interest. According to [1], Montano et. al. reviewed and analyzed the state of the art regarding triage applications for health emergencies. Based on their systematic review of the literature in the scientific database from 2010 to early 2021, only 13 applications were identified from 26 relevant papers. Surprisingly, they also observed that despite the existence of much research, only 3 applications are accessible.

Usually, the triage are guidelines in a paper form; for example, clinical guidelines for major incidents and mass casualty events proposed by NHS England [2]. In the UK, triage at mass casualty incidents is performed in two steps. The first step takes place at the scene of the incident.

The primary triage assessment takes no more than 30 seconds per patient. Its objective is to rapidly identify those patients who need a life-saving intervention. According to [3], the first step is executed by using an algorithm such as the Modified Triage Sieve – MPTT-24.

The second step takes place in a more permissive environment and is performed by a more experienced clinician. There are more types of the secondary triage processes, such as MIMMS Triage Sort, Anatomical Triage, and Clinical Gestalt. For our research, we used the algorithm that was proposed by our colleagues [4].

3 Tools

The tools we focus in this paper are a mobile application designed for triaging in the field and a triage data management platform, where the data acquired in the field is stored and analyzed by specialists. Primarily, the mobile app stores and shares the data locally with all medics and paramedics in the field via Bluetooth LAN as there might be problems with internet connection (see Fig.1). When there is internet connection, the triage data is synchronized with the server and stored into the server database and further managed from the web platform. In the next subsections we describe these tools.

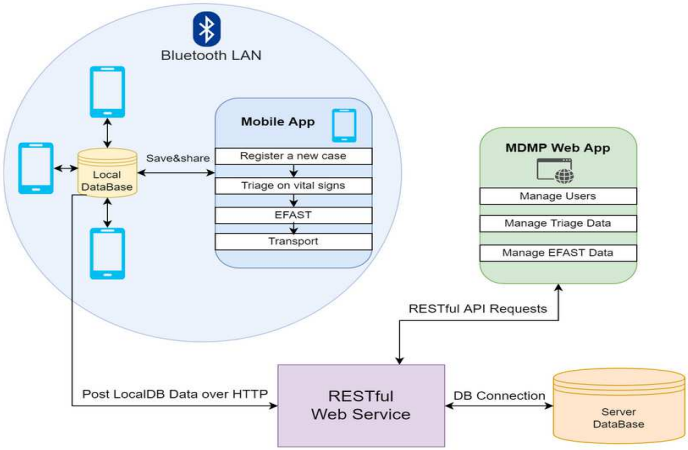


Figure 1. Architecture of communication between tools

4 Mobile Application

Mobile Application was developed by using Android Studio Arctic Fox v.2020.3.1. powered by the InteliJ Platform (see Fig.2 a.).

National Association of Emergency Medical Services Physicians (NAEMSP) proposed SALT [5] to triage and move Mass Casualty Incidents (MCI) patients forward to resources. SALT is a four-step process for first responders to manage mass casualty incidents, and stands for:

1. **Sort**;
2. **Assess**;
3. **Lifesaving interventions**;
4. **Treatment and/or transport**.

We define 3 steps in our triaging process:

1. **FAST**;
2. **EFAST**;
3. **Transportation**.

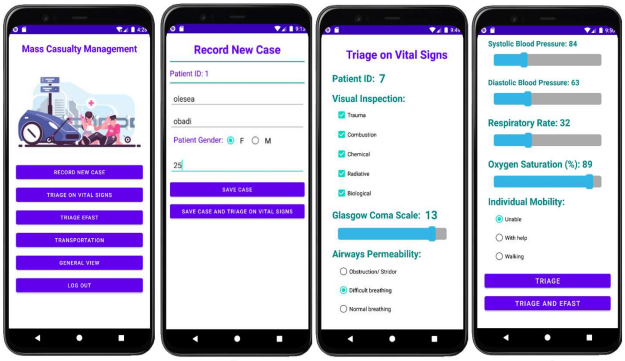


Figure 2. Mobile application home interface and FAST process interfaces

The FAST process, also called the triage on vital signs, has the feature to be completed in one go by one medical assistant (see Fig.2 b.-d.) or step-by-step by 2 people. For the cases, when they are completed by 2 people, one of them completes only **Record New Case** interface (see Fig.2 b.), the other one completes **Triage on Vital Signs** interface (see Fig.2 c., d.).

Our colleagues from Informational Systems Laboratory [6] developed a scoring system based on decision rules that allows re-assessment of triage priorities for casualties. Structurally, the score is in line with the well-known scoring systems, widely used in medical diagnostics. Based on a scoring system algorithm, our application at the end of completing the triage, on vital signs gives a quick categorization of casualties in Red I, Red II, Yellow, and Green, (see Fig.3 a.). The Red categories need immediate transportation, while the Green one is processed the last in order.

The added general view window allows users to visualize all basic information about injured persons: ID, First name, Last name, Gender, Age, Patient Status (Non-triaged, Triaged), and Triage category (see Fig.3 b.). Selecting any casualty from this list allows the user (with the

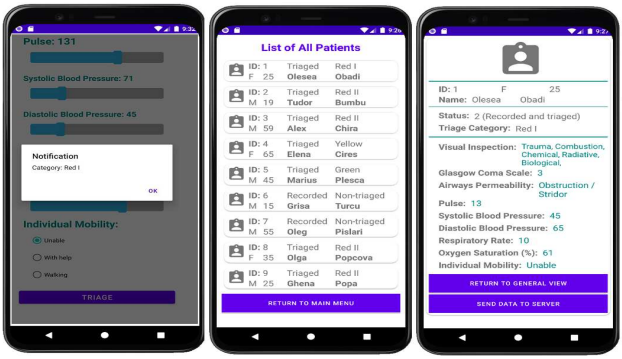


Figure 3. Notification category based on system score algorithm and General View interfaces

corresponding rights) to proceed to the next stage of the management of mass casualty situations, or to monitor the overall performance (see Fig.3 c.).

The stage of triage EFAST (see Figure 4.a) allows to introduce information, using additional portable ultrasound equipment, regarding presence of free fluid in the following areas:

1. Right upper quadrant
2. Left upper quadrant
3. Pelvic view
4. Pneumothorax (left and right)
5. Hemothorax (left and right)
6. Cardiac tamponade.

The decision rules were created for EFAST, helping to clarify the triage category more accurately, for instance, to change it from Yellow into Red in case of free fluid detected (presumed to be blood under disaster conditions). Medical experts have validated these decision rules.

The stage of transportation allows medical assistants to select the corresponding type of ambulance (type C – intensive care ambu-

lance, type B – emergency ambulance, type A – non-emergency transportation) from the available ones, destination hospital, and the order of transportation (see Fig.4 b.,c.).

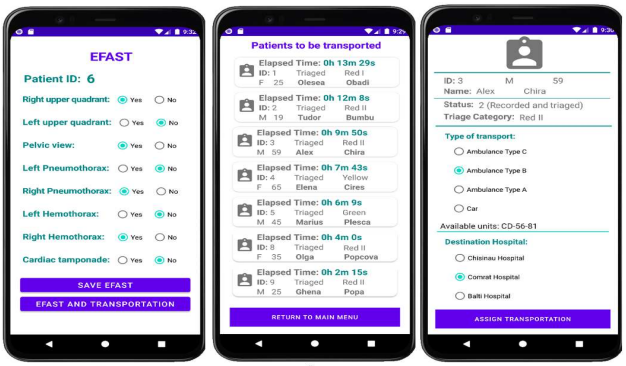


Figure 4. a. The EFAST process interface; b. and c. The transportation process interface

For every stage of the management of mass casualty situations, there was tested the stage's main functionality, modules integration, and if the developed software meets the needs of the user. While designing interfaces for mobile applications, the factor that medical assistants will use gloves was considered, so all buttons should be as big as possible. The main factor in designing was that the time consumption should be as short as possible. Therefore, the interface is intuitive and friendly.

5 Medical Data Management Platform

Data collected in the mobile application during the stage of triage of vital signs can be saved, stored, and managed from a web application developed within the Django framework. We call this project Medical Data Management Platform or MDMP (see Fig.5).

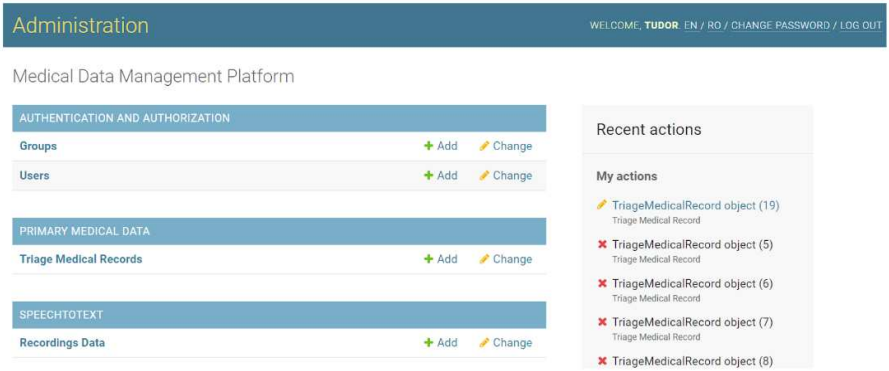


Figure 5. The MDMP homepage in English

The MDMP consists of 3 modules: the User Group Administration module; the Patient Medical Data Administration module; and a special module for audio processing and management, namely the Audio Recordings Administration module which is still in development and not described in this paper. The platform communicates with the mobile application via REST API by requesting the data of registered patients triaged as long as internet connection is established. The API allows GET and POST request methods.

In the User Group Administration module, the site administrators can manage the groups and users of the platform (see Fig.6). Also, the administrator can manage the permission of a user or a group of users, such as view, edit, or delete medical records. The following groups of users were added: Triage Members, the Medical Personal consisting of medical assistants and doctors which will review the data in the Primary Medical Data Administration module and prepare reports. Also, there is the group of Administrators including the managers and MDMP administrators. In this module, the Administrator can add new groups of users.

The next module is Patient Medical Data Administration (see Fig.7) for managing recordings of registered and triaged patients sent

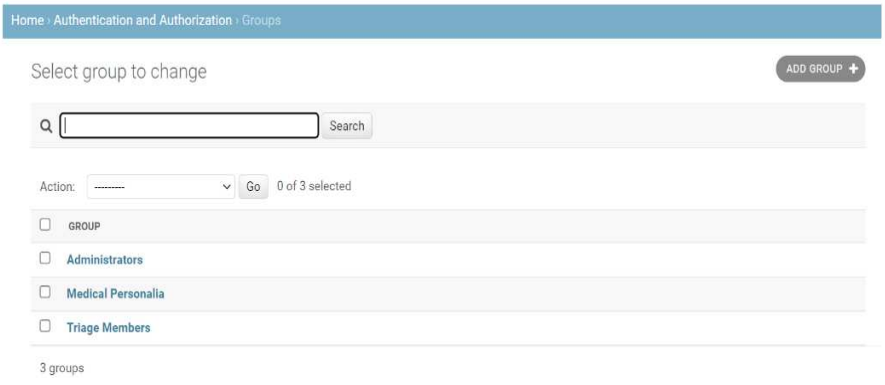


Figure 6. User groups management page

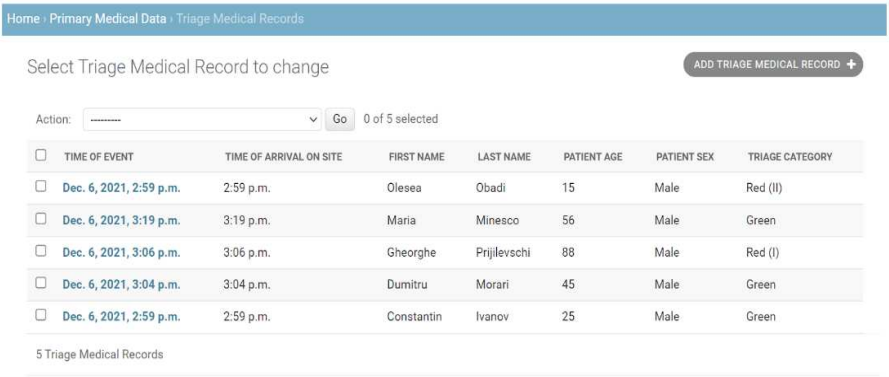


Figure 7. Patient medical data administration pag

from mobile application. The users within the Medical Personal group can view, edit, or delete these recordings. A recording from the table shown in the Figure 7 displays 7 attributes which include one of the most important attributes which is the Triage Category.

6 Conclusion

The developed tools, i.e., the mobile application for triage and the web platform for managing and analyzing casualties offer simple and user-friendly interface, allowing medical first aid personnel the following actions: to gather and organize the primary medical data of casualties; to perform triage based on vital signs and assign the triage priority for quick categorization of casualties **(Red I, Red II, Yellow, Green)**; to store and analyze the data acquired in the field.

It can be done for casualties with injuries and can be repeated during transportation; the priorities can be set for evacuation of injured persons from the disaster site and for routing them to the specialized medical centers (including the transmission of the casualty-related data).

The impact of these tools is crucial by doing fast registration and triage priority assessment; accurate casualty triage reassessment and more effective emergency therapy before further transportation, that will minimize over- and undertriage; coordinated evacuation of casualties will help in efficient distribution of the available resources; data analysis for improving the mass casualty management.

Acknowledgments. "Intelligent Information systems for solving ill structured problems, knowledge and Big Data processing project" Ref. Nr. 20.80009.5007.22, has supported part of the research for this paper.

References

[1] I.H. Montano, I.T. Diez, R.L. Izquierdo, M.A. Castro Villamor, and F.M. Rodriguez. *Mobile Triage Ap-*

plications: A Systematic Review in Literature and Play Store. J Med Syst, 2021; 45(9): 86. Published online 2021 Aug 13. doi: 10.1007/s10916-021-01763-2. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8361243/.

[2] NHS England (2019). *Clinical Guidelines for Major Incidents and Mass Casualty Events.* Guidelines online. Available: https://naru.org.uk/wp-content/uploads/2019/01/version1__Major_Incident_and_Mass_casualty_guidelines-Nov-2018.pdf. [Accessed 25 Jul 2022].

[3] R. Mersh and J. Vassallo. *Triage in Mass Casualty Situations* RCEM Learning. Published online: 06/01/2020. Available: https://www.rcemlearning.co.uk/reference/triage-in-mass-casualty-situations/#1572966037676-482ad5bd-c7b3. [Accessed 12 Aug 2022].

[4] S. Cojocaru, C. Gaindric, I. Secrieru, S. Puiu, and O. Popcova. *Multilayered Knowledge Base for Triage Task in Mass Casualty Situations*. Computer Science Journal of Moldova, vol. 24, no. 2(71), 2016, pp. 202–212.

[5] R. Duckworth. *How to use SALT to triage MCI patients*. Published on EMS 1 by LEXIPOL, 27 April, 2021. [Accessed 9 Aug 2022]. Available: https://www.ems1.com/mass-casualty-incidents-mci/articles/how-to-use-salt-to-triage-mci-patients-ioh8pD88282FDTdy/.

[6] Iulian Secrieru, Constantin Gaindric, Elena Guțuleac, Olga Popcova, Tudor Bumbu. *Formalization of decision knowledge and reasoning for casualty prioritizing*. Proceedings of the Workshop on Intelligent Information Systems (WIIS2022), to be published.

Olesea Caftanatov[1], Tudor Bumbu[2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: olesea.caftanatov@math.md

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: tudor.bumbu@math.md

Formalization of decision knowledge and reasoning for casualty prioritizing

Iulian Secrieru, Constantin Gaindric, Elena Guţuleac,
Olga Popcova, Tudor Bumbu

Abstract

Primary triage, based on vital signs, is the first and the most important stage of mass casualty situations management. There is still no single approach, both among experts in emergency medical care and among developers of decision support systems. This is explained by the existence of various medical protocols (mandatory for execution at the national level), focusing on the main types of disasters that prevail in the region.

In this article we describe the formalization process, aiming to create an efficient inference for casualty prioritizing, based on vital signs.

Keywords: medical informatics, pre-hospital triage, mass casualty situations, knowledge acquisition and formalization.

1. Introduction

Knowledge formalization and the corresponding reasoning schemas selection is considered the key phase in the development of any inference module.

The domain of mass casualty situations and disasters management, on the one hand, is well structured, on the other hand, there is no general and universal approach in the world. The same applies to the main stage of the process of management – casualties primary triage, based on vital signs (see Fig. 1 - Obj. 1). Different countries design their triage model for emergencies according to their native medical protocols, resources, and forces.

The main drawback of many of the proposed approaches is that at the knowledge acquisition stage, given that national protocols/standards are 'over-structured', and the developers are forced to choose 'rigid' schemas of its representation. As a result, the inference does not correspond to the daily work and habits of the first-aid person, who is the end-user.

The discrepancy between the inference module of the knowledge-based system and the form of the doctor's diagnostic reasoning may become the cause of different mistakes or it may lead to rejection of the user to utilize it in medical practice.

Decision knowledge and reasoning formalization techniques, aiming to create an efficient inference for casualty prioritizing, based on vital signs, are described in this article.

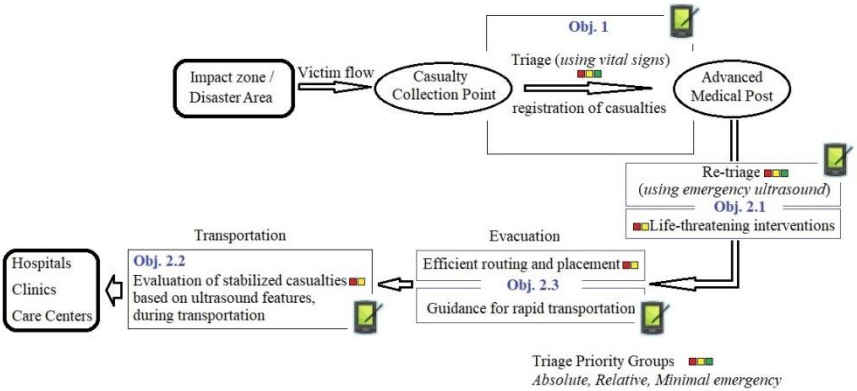


Figure 1. Victim flow management. Example scenario

2. Knowledge representation schemas

The main goal of the knowledge representation schema is to represent professional knowledge in a manner to facilitate drawing conclusions (inferencing) [1].

We distinguish two approaches: single and hybrid knowledge representation schemas.

Hybrid schemas represent integrations of two or more single knowledge representation schemas.

The most popular single knowledge representation schemas are the following:

- Decision trees and their descendants (frames or schemas);
- Semantic nets, Conceptual graphs, ontologies;
- Symbolic rules;
- Fuzzy rules (fuzzy logic);
- Case-based representations;
- Neural networks;
- Belief networks (or probabilistic nets).

Semantic nets, decision trees, and ontologies represent knowledge in the form of a graph (or a hierarchy) [2]. Nodes in the graph represent the concepts and edges represent the relations between the concepts. All of these knowledge representation schemas are very natural and well suitable for representing structural and relational knowledge.

Symbolic rules (symbolic reasoning) are one of the most popular knowledge representation schemas [3], representing general domain knowledge in the form of IF-THEN rules:

if <conditions> then <conclusion>,

where the term <conditions> represents the conditions of a rule, whereas the term <conclusion> represents its conclusion. The inference engine uses the knowledge in the rule base as well as facts about the problem at hand to draw conclusions. The efficiency of the inference process depends on the length of the inference chains.

3. Casualty prioritizing in mass casualty situations

By studying the existing types of medical triage, we can distinguish two main logical approaches: algorithmic [4] and numerical [5].

When the algorithmic model is used, the casualty should be assigned to a certain category (casualties with severe, moderate, or minor injuries), taking into account every vital sign (ability to move, breathe, level of consciousness, etc.). If the considered parameter is within the normal range, then the next parameter is studied according to its priority within the triage system.

When the numerical approach is used, the doctor who is performing medical triage should simultaneously assess all parameters of the model. As a result, the final assessment of the casualty state is made, being based

on the overall assessment of all parameters of the model. In accordance with the final assessment, the casualty is assigned to one of the categories of this medical triage.

4. Formalization of decision knowledge and reasoning. Knowledge acquisition

In collaboration with a team of medical experts in clinical emergency medicine, the minimum set of parameters needed for casualty registration was identified, so that the record, accompanying the casualty, contains all the information that will enable doctors from specialized medical centers to intervene operatively in the treatment.

These parameters cover all stages of the initial assessment of casualty and the organization/structuring of primary medical data.

The medical record for casualty registration consists of personal data, time interval, type of injury (resulting from visual inspection of casualty), and values of basic attributes (parameters) that describe vital signs.

The following 9 parameters were selected: visual inspection, Glasgow Coma Scale, airways, pulse, systolic and diastolic blood pressure, respiratory rate, oxygen saturation, and individual mobility. As the knowledge representation schema, there was selected the tabular form (most often used in the field of emergency medicine) [6].

There was developed an acquisition web-module – Medical Data Management Platform (https://g5700.math.md/admin/).

This platform allows one to record and store data about the casualty state, prioritizing them in 4 emergency categories (RED (I), RED (II), YELLOW, GREEN), based on vital signs.

5. Knowledge base kernel creation

The inference module is based on 4 decision rules, identified in collaboration with medical experts.

As the formalization schema of the decision rules, there was used symbolic rules, once again – being one of the most popular methods in the field of medical information systems.

All 4 rules were formalized and integrated into the inference module, representing now the knowledge base kernel.

Find below how a rule is represented in the inference module.

Pseudo-code (Java):

```
// conditions for category Red I
if ((glassgow_coma >=3 && glassgow_coma <= 8) &&
      (airways==1) &&
      (pulse<40 || pulse > 120) &&
      (systolic<80) &&
      (respiratory_rate<13 || respiratory_rate > 35) &&
      (oxygen <= 85) &&
      (mobilitiy==1)) {
   alert("Category: Red I");
   triage_category = 1;
   return triage_category;
}
```

These rules represent the scoring system for the triage of casualties. To validate this result, it was decided to create a synthetic data set and pre-test the decision algorithm.

6. Pre-testing and conclusion

The pre-testing process showed: i) a large number of cases with an 'undetermined' emergency category, if a totally random selection is applied.

The process of creating the data set showed: ii) the need to use a systemic approach to generate the synthetic data set; iii) the need to create an additional field in the medical record for a casualty on the web platform – for the comments of medical experts in the synthetic data validation process.

As a result, a synthetic data set of 56 cases (medical records) was created, following the approach below:

A) Values for all 9 parameters were selected from the same column, representing some emergency category;

B) 7-8 values from one column and 1-2 values from the neighboring column(s).

For 32% of these 56 cases, the emergency category was determined unequivocally, for 68% – the opinion of medical experts is required. In this sense, there was proposed and implemented the creation of an

additional field in the medical record for a casualty on the web platform – for the comments of medical experts in the synthetic data validation process.

The obtained result allows the authors to state that the proposed approach is a viable one for creating an efficient inference for casualty prioritizing, based on vital signs. For the selection of all used methods and technologies, there were formulated the grounds and rationale.

7. Future work

Both the study of the problem domain and the obtained result showed the existence of a one-sided orientation in the selection of the way to describe the reasoning for casualty prioritizing – algorithmic (decision trees, decision rules, etc.) or numerical (tabular form, scoring system, etc.).

At first glance, this selection is determined by the type of data source and the data itself. If developers have access to expert data and experts, which can formulate their professional knowledge in the form of rules, then the algorithmic approach is chosen. Otherwise, if developers have access only to the set of precedents (real cases) – the numerical approach is chosen.

Another reason is the time available for decision-making. If time is extremely limited, then the algorithmic method is chosen because it allows one to generate the conclusion without specifying the values for all parameters. If time allows for determining all the values for the system of selected parameters, then the numerical method is chosen, and even more, an attempt to create a scoring system is made.

A deeper analysis of the obtained result allowed the authors to make the following hypothesis: consciously or unconsciously the developers of medical information systems make their choice based on the restrictions imposed by the end-user habits – first-aid person and/or by the subdomain, in which the future information system will be used (disaster type like an earthquake, chemical or nuclear accident, explosion, flood, etc.; restrictions by age or any anthropometric data).

To verify the formulated hypothesis, it would be interesting to carry out both approaches – algorithmic and numerical (including a scoring system) for the same restrictions of the problem subdomain and the same set of real cases, described with the same set of 9 parameters.

Acknowledgments. The Moldovan State Program project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" has supported part of the research for this paper.

References

[1] Iu. Secrieru. *Structured knowledge management techniques for the development of interactive and adaptive decision support system*. In: Computer Science Journal of Moldova, nr. 1 (49), vol. 17, 2009, Kishinev, pp. 58-73, ISSN 1561-4042.

[2] M. Negnevitsky. *Artificial intelligence: A guide to intelligent systems*, Reading, MA: Addison Wesley (2002).

[3] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, (2003), ISBN 0-13-790395-2.

[4] Gh. Ciobanu, M. Pîsla, F. Gornea et al. *Ghid naţional privind tirajul medical în incidente soldate cu victime multiple şi dezastre*. Centrul Nat. Șt.-Practic Medicină de Urgenţă, Centrul Republican Medicină Calamităţilor. – Chişinău, 2010, 36 p, ISBN 978-9975-80-337-3.

[5] S.P. Gormican. *CRAMS scale: field triage of trauma victims.* Ann Emerg Med. 1982 Mar; 11(3), pp. 132-135.

[6] C. Gaindric, S. Şandru, S. Puiu, O. Popcova, Iu. Secrieru, and E. Guţuleac. *Advanced pre-hospital triage based on vital signs in mass casualty situations.* Proceedings of the Workshop on Intelligent Information Systems WIIS2021, October 14-15, 2021, Chisinau, Republic of Moldova, pp. 130-134, ISBN 978-9975-68-438-5.

Iulian Secrieru[1], Constantin Gaindric[2], Elena Guţuleac[3], Olga Popcova[4], Tudor Bumbu[5]

[12345]Affiliation: Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chisinau
E-mails: [1]iulian.secrieru@math.md, [2]constantin.gaindric@math.md [3]elena.gutuleac@math.md, [4]oleapopcova@yahoo.com, [5]tudor.bumbu@math.md

Part 4

Automatic content generation systems for computer-assisted training

Important aspects in assessing the credibility of unstructured information

Mircea Petic, Adela Gorea, Ina Ciobanu

Abstract

In this paper we presented possibilities of credibility assessment tools. A large number of such tools have been developed for journalists. Another important subject of the paper is the actual approaches that are used in the process of assessment of credibility of unstructured information. In this sense, two points of view were presented.

Keywords: credibility, Web, social networks, fake news.

1. Introduction

With the development of online media, since the advent of the Internet, considered a democratic space, it has become much easier for anyone to express themselves freely, anytime and anyway. Beyond the media institutions, each with its own editorial policies and various frameworks that judge whether an event turns into news or not, there have been many sites that present themselves online as media products, posting various content that puts as doubt its credibility. Whenever we are about to accept or reject new information, we should ask ourselves what is the origin and reputation of the source. In the so-called "reputation era", critical appraisals should be directed not to the content of information, but rather to the network of social relations that has formed that content and offered to it a certain position, deserved or undeserved, in our knowledge system. But not every user is able to make an analysis and a distribution of content into a credible or untrusted one.

Social media networks also play an important role, because through them the content is distributed and redistributed daily by users, without being checked, often creating feelings of panic and revolt.

The issue of assessment of the credibility of unstructured information is one of the most important and is not just specific to digital media. Finding credible data is a task of great interest to those seeking information. The information is quite difficult to process due to its unstructured form. Even credibility coalition[1] was created which aims to understand the veracity, quality and credibility of online information.

The aim of this paper is to underline the possibilities of the credibility assessment tools and to detect the important approaches that are used in the process of assessment of the credibility of unstructured information.

That is why the paper is structured mainly in two parts. The first part is concerned to the tools for assessing information credibility. A large number of such tools have been developed for journalists. The second part is dedicated to the research methods for assessing data credibility. In this sense, two points of view are presented.

2. Tools for assessing the credibility of information

Several existing tools that assess the credibility of online texts and / or articles are analyzed, focusing mainly on user-generated evaluations as experts in journalism.

FactCheck.org is an information verification platform that was launched in December 2003. On this site, users can ask questions that are usually based on a rumor in politicians' statements. The site team conducts an investigation and provides a detailed explanation. The explanation includes information about who is the author of the statement, when it was released and how the team verified it. The site also has a special function for verifying scientific information.

Politifact.com is an information verification platform and one of the first fact-checking newsrooms in the US, founded in 2007. The group of reporters within this platform monitors the statements and speeches of politicians and denies false information [1].

[1] https://credibilitycoalition.org/

Snopes.com is an information verification platform developed in 1994 and aims to validate statements, articles, posts, photos on social media. This platform is not limited to simple statements (eg "true" or "false"), but uses more detailed categories ("true", "false", "partly true, partly false", "largely true"). , "Mostly false", "outdated information", "misunderstood information", etc.).

Fake Bananas is a tool developed by a group of Sworthmore College students. The tool is based on machine learning algorithms and defines credibility with 82% accuracy. The program searches for authorized online publications for articles with the context of the message, which must be verified and analyzes whether the authors of the articles agree with the formulated idea made in the statement. If trusted sources agree with it, the program evaluates the statement as true. Although the service is not publicly available, the program can be used in other projects [1].

Hoaxy is a tool developed in 2016 by a group of researchers at the Center for Complex Networks and Systems Research and the Indiana University Network Science Institute. The tool was developed to study how information is disseminated on social media. Focused on checking for fake news, the site generates interactive, color graphics so that users can see how messages are spread on Twitter.

NewsGuard is a tool dedicated to source-level evaluation, manually and methodically reviewing thousands of English-language news sources, mainly in the US. NewsGuard is available as a Chrome extension that can display this information when such news sources are open in the browser or appear in some web searches. These criteria are divided into two groups: credibility (does not repeatedly publish false content, regularly corrects or clarifies errors, collects and presents information responsibly, avoids misleading headlines, responsibly manages the difference between news and opinion) and transparency (site discloses ownership and funding, discloses who is responsible, clearly labels advertising, site provides name of content creators, along with contact information or biographical information) [2]. Websites also receive an overall score (the sum of points for each criterion) and a tag, which can be reliable, negative, satirical or platform (blogs, user-generated content or social networks). The total score of credibility and transparency is a maximum

of 100, and a news site is considered "safe" if it has accumulated at least 60 points.

My Web of Trust (WOT) is a reputable crowdsourced service that provides evaluations of websites through a browser extension. It offers two components - trust ("How much do you trust this site?") and security ("How appropriate is this children's site?") - in terms of a score and a measure of trust. Users can view general ratings and comments from the community and provide their own ratings. The user can also rate it and leave a review based on his personal impressions. WOT has two modes: real-time protection and manual mode. Real-time protection informs you about online threats [3].

If you find a site that does not yet have a reputation rating, you can ask the WOT community to rate that site. For ratings and reviews, WOT uses smart algorithms and manual verification to detect and remove fake reviews and can also use it to check blogs and social networks (Twitter, Facebook, Google+).

Although a significant number of scientific papers address many aspects of this topic, few have researched methods of measuring data credibility. There is almost no research that would propose algorithms for assessing the credibility of content that would allow the automation of solving this problem to an extent acceptable (accurate and useful) by users.

3. Research methods for assessing data credibility

According to [4], but also according to the analyzed tools from above, credibility assessment methods can be divided into three broad categories: methods based on automated approaches, methods based on human assessment and mixed methods, which in turn are divided into other subcategories. The combined method unifies main categories or different subcategories.

Speaking about automated-based approaches concerning credibility assessing we should speak about machine learning approaches, graph-based semi-supervised approaches and weighted algorithm and information retrieval algorithms.

Human based approaches consist of cognitive and perception approaches, voting approaches and manual verification approach.

Mixed methods use advantages of the post, topic or user level credibility assessments.

Moreover there is also another classification of the methods in credibility assessment of online information that is based on approaches related to social networks and linguistics approaches [5].

If we talk about the approaches related to social networks, then we take into account the notion of linked data[2] and the behavior of users on social networks. Just as users need to log in before using a social network, it gives them more confidence in the data that appears here.

In the case of computational linguistics approaches, we speak about the use of statistics on n-grams, the sentences transformation into more advanced forms of information representation (such as decision trees), to which the attached probabilities are then analyzed to identify anomalies, semantic analysis of information, relationships between linguistic elements, which help to determine the proximity to the centers of truth or deception, SVM classifiers or Bayesian Naïve classifiers are used to predict the outcome, use neural networks that identify false news.

However, the hybrid approach (combining machine learning in computational linguistics with social networking approaches) seems reasonable and promising [5].

5. Conclusion

Analyzing the web tools mentioned in the paper, we see a continuous effort to assess the credibility of information on the Internet. The use of the tools described in the paper confirms that the correct classification of the multitude of information is not a simple activity. Even if there are several applications, they cannot guarantee accurate results. Moreover, the results are modest for inhomogeneous documents that have a complex structure that contain not only texts but also images. Most of the researched applications work only for English, which proves that it is practically an unexplored field.

[2] https://www.w3.org/standards/semanticweb/data

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”[3].

References

[1] S. Molly *8 resources for detecting mis- and disinformation*, 2019, International jurnalists network, https://ijnet.org/en/story/8-resources-detecting-mis-and-disinformation.

[2] S. Wineburg, S. Mcgrew, J. Breakstone, and T. Ortega. *Evaluating information: The cornerstone of civic online reasoning*. In: Stanford Digital Repository. Retrieved January, 8:2018. DOI: 10.1145/3170427.3174355.

[3] F. Prochazka and W. Schweiger. *How to measure generalized trust in news media? An adaptation and test of scales.* In: Communication Methods and Measures, 13(1):26–42. DOI: 10.1080/19312458.2018.1506021.

[4] M. Alrubaian and M. Al-Qureshi. *Credibility in Online Social Networks: A Survey*. In: Digital Object Identifier 10.1109/ACCESS.2018.2886314. 2028-2055 pp. 2018. DOI:10.1109/ACCESS.2018.2886314.

[5] A. Iftene. *Exploiting Social Networks. Technological Trends*. In: Habilitation Thesis submitted at "Alexandru Ioan Cuza" University, December 2019. 163 p. ISSN 2668-1765, ISSN-L 1224-9327.

Mircea Petic[2,1], Adela Gorea[1], Ina Ciobanu[1]

[1] Alecu Russo Balti State University, Republic of Moldova
E-mail: adela.gorea@usarb.md, ina.ciobanu@usarb.md

[2] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Republic of Moldova
E-mail: mircea.petic@math.md

[3] http://www.math.md/en/projects/20.80009.5007.22/

Generation and use of educational content within adaptive learning

Alexandr Parahonco, Mircea Petic

Abstract

One of the new educational technologies that has shown its undoubted effectiveness is e-learning. Nowadays, beyond e-learning, adaptive platforms have been appeared with commercial services for any sort of adaptation. In the Republic of Moldova where universities are not highly financed, such commercial systems are unprofitable in usage. Our solution lays under development of our own adaptive learning system on the base of the Moodle Web platform.

Keywords: adaptive learning, plug-in, crawler, Flexible, TestWidTheory, TestWid.

1 Introduction

Intelligent adaptive learning systems originate fast, but still incur obstacles in realization. Theoretically these systems must organize learning process based on differentiation for each student. Formative and diagnostic assessment should apply adaptive intelligence for precise results. The problems arise in the ways of implementation of these principles and involved technologies.

Thus, in the 21st century, the idea of adaptive learning becomes even more popular: not only teachers and psychologists, but also managers and businessmen are interested in it. Subsequently, such popular adaptive educational systems as Knewton Alta, Smart Sparrow, Geekie, ALEKS and others appeared. Such systems to a greater extent

are highly specialized (to the domain of studies, for example) and paid. That is why higher education institutions from Republic of Moldova cannot use them. Our universities need free platform without any domain limitation.

The purpose of the article is to study the principles of generating educational content and the development of an adaptive learning system for higher education institutions of Republic of Moldova that will be capable to compete against leading commercial products.

The article begins with an analysis of the systems of adaptive learning, the development of their common models of adaptive learning and an assessment of their relevance to the usage in the educational process of the university. Then it discusses the developing model for Moodle Web platform as the base for new adaptive learning system. At the end of the article, practical solutions are proposed for the implementation of adaptive learning in the framework of the national system of higher education [1, p. 163].

2 Analysis of the existed systems

Adaptive learning systems represent educational information and communication technologies that respond in real time to student actions and, in accordance with the information received, provide him with individual support [2, p. 8].

When creating an adaptive educational system, first of all, three key questions are solved: what is modeled, how it is modeled and how the adaptation model is supported. Then one of three scenarios is implemented, where the adaptation object can be: **content, tasks or the order of presentation** of educational materials [2, p. 9].

If in the educational system the **object of adaptation** is content, then it functions according to the following algorithm: first, it analyzes the student's response to the task and, in the case of an incorrect answer, offers him feedback, tips, or additional educational materials. For example, Geekie is a paid learning platform powered by artificial intelligence (AI) to prepare Brazilian students for their final exams. This

platform provides adaptation at the level of curriculum modification [3].

Smart Sparrow offers **three levels of customization**: feedback, curriculum modification (learning paths), and the ability of the educator to facilitate the transmission of knowledge. Functionality of this platform allows teachers to create interactive content that can be adapted for any group of students according to the topic of the subject under study and the specific requirements for the group's learning process. Teachers can determine individual learning paths for students, interact with them in real time, and also use a number of ready-made templates to save time when creating electronic content [4, p. 374].

Knewton Alta platform is well known for its programs and applications with adaptive functions. Knewton's team of specialists managed to create universal algorithms and develop an extensive infrastructure for collecting, analyzing and using information about student progress. It contains **two elements of adaptation**: tasks and the procedure for providing materials [5, pp. 202-207].

One of the developers of the adaptive tests for monitoring is NWEA (Northwest Evaluation Association) that creates the adaptive tests for different goals. For example, the test MAP Growth is used for the periodic testing of pupils' knowledge of different subjects, while MAP Skills is recommended to be applied more often [6, pp.553-554].

All these educational platforms – Knewton Alta, Geekie, Smart Sparrow and NWEA – are not suitable for the university's higher education system, because they are expensive, overly flexible, which is inconsistent with the structure and time constraints of courses; are time consuming.

To summarize, we propose to create new adaptive learning system on the base of Moodle Web platform. Moodle competes on an equal footing with the world flagships of the distance education system market. Thanks to this, Moodle combines a wealth of functionality, flexibility, reliability and ease of use. That is why, most of our universities use this educational platform what played an important role in our decision.

3 Moodle model of adaptive learning. Adaptive test

Due to the fact that it is a "plug-in base system", we developed at first model for adaptive learning. To form an adaptive course, we offer a model (Fig. 1), which consists of three separate plugins: a plugin for adaptive testing **TestWid**, a plugin for storing educational materials (text, video, audio, images, exercises, tests) **TestWidTheory**, and a plugin for an adaptive course **Flexible**. The degree of adaptation should cover all three levels: content, order of presentation and tasks.

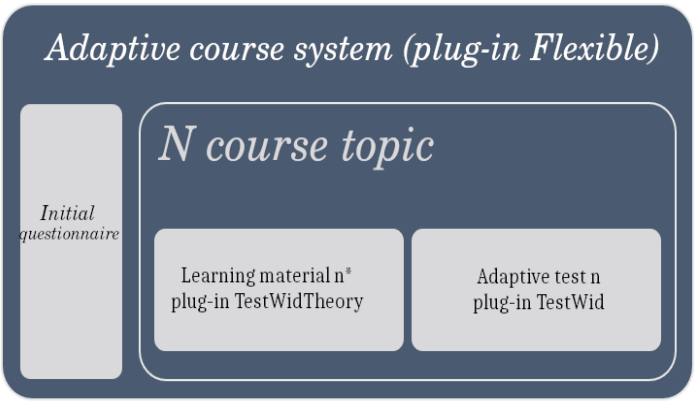


Figure 1. Adaptive course model for the Moodle platform.

This system should allow the teacher to generate course content and supplement it for each student, based on his preferences. These preferences will be determined in the **initial survey** (questionnaire) with 2 simple questions with the list of answers: 1) sort in descending order types of learning materials according to your opinion (videos, images, text, learning games); 2) sort in descending order the source list for custom content search (google, youtube, wikipedia, encyclopedias, others).

When saving this data, the system (Flexible plug-in) will arrange

learning resources in each topic by priority.

The developing system requires that the assessment of the quality of students' knowledge and their competencies is determined using an adaptive test (TestWid) for each topic. It is not prohibited to use other plugins (test, assignment and others) to control the quality of training, but the assessment will not be considered the main one. This rule also applies to plugins for learning resources: at least one instance of the TestWidTheory plugin must be in each theme. The idea of using these two plugins together is dictated by the connection of each piece of theory with a specific question from the test. So, a teacher primarily creates the content, then selects by the mouse some fragments and, clicking the button "Adaugă teoria la TestWid", selects from drop-down list from which adaptive test and towards which question to link the fragments.

Thus, if a student has not formed the required set of competencies and has not passed the test, he will be presented with a list of tasks with incorrect answers and a link to related fragments of the theory (Fig. 2, 3).

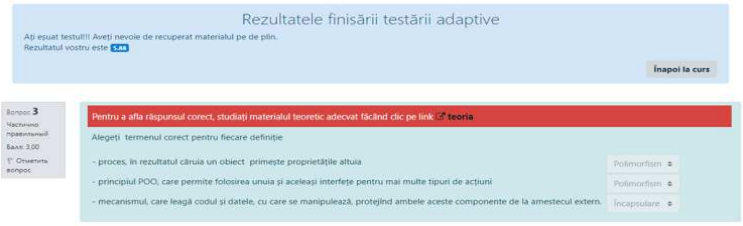


Figure 2. List of incorrectly completed tasks – question number 3.

After repeated self-preparation, the test will be retaken. When retaking, the testing system will select questions from each category that are different from those already used, if any.

The development of an adaptive testing model for the TestWid plugin was based on the **"Three-level algorithm"** model [7, pp. 233-234], which takes into account the capabilities of the Moodle platform

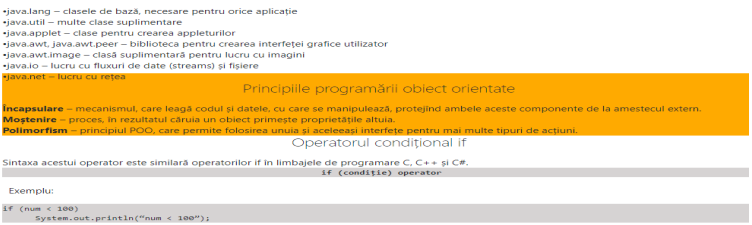


Figure 3. Fragment of theory related to question 3 after following the link.

(fixed number of questions in the test). The "three-level algorithm" allows, in the presence of 15 questions-tasks, to achieve the same accuracy and reliability as in the test with 45 exercises that do not pay attention to their level of complexity, and also allows three times to reduce the cost of testing duration, while maintaining information security.

Therefore, the question with the **adaptive path** of the student is solved. Against this background the question arises about both answer assessment and formula of the final grade. The solution comes from the **mathematical model** for assessing knowledge based on **learning levels**. The characteristic of an assignment is the level of assimilation, for which it is intended to test. Tasks can be divided into five groups corresponding to the levels of assimilation: understanding, identification, reproduction, application, creative activity [8, p. 12-13]. A set of essential operations is determined for each task. Essential operations are those operations that are performed at a verified level.

Thus, to assess students' answers and knowledge, the coefficient K_a is used (1):

$$K_a = \frac{P_1}{P_2}, \tag{1}$$

where P_1 – the number of correctly performed essential operations in the control process;

P_2 – the total number of significant operations in the test;

$\alpha = 0, 1, 2, 3, 4$ – corresponds to the levels of assimilation.

The grade is set on the basis of the specified cut-off values by ratios multiplied by 10:

$K_a < 0.7$ – unsatisfactory

$0.7 \leq K_a < 0.8$ – satisfactory

$0.8 \leq K_a < 0.9$ – good

$K_a \geq 0.9$ – excellent.

Finally, due to the "three-level algorithm" and level of assimilation it is possible to create truly adaptive test that can estimate student's knowledge and skills.

4 Model for the dynamic content generation for training courses

According to the object of adaptation, content generation should allow users (teachers and students) to generate learning content. Such kind of system can be developed by the **web-scraping** technology, including **Data mining** and **text mining** approaches. The thorough study of technical documentations, science articles and practical experience, we designed the Scheme of the program model for the dynamic creation of e-courses (see Fig. 4).

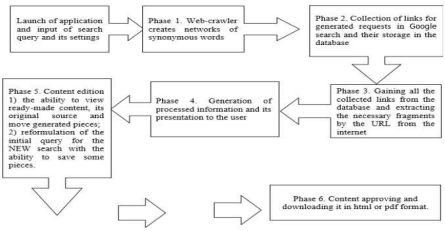


Figure 4. Scheme of the program model for the dynamic creation of e-courses.

As it can be seen, at Phase 1, the operating principle of the developed model is to use synonymous connections to search for dictionaries

that are similar in meaning with the help of a crawler, and use them at Phase 2 and 3 for advanced search using the Google search service. Thus, we obtain the behavior model of a user performing manual scraping.

According to phase 4 and 5, well merged content should be generated and further exported in Moodle. The content may be imported in Moodle via Page and File standard plugins at Phase 6.

Our application will allow editing content (font, placement, color) and downloading the files in html or pdf formats.

5 Description of the design and principle of operation of the Flexible plug-in

All plugins developed within the flexible course were created based on the documentation of the Moodle developers [7] and are designed to work on Moodle platforms starting from version 3.

The work of any plugin on the Moodle platform begins with its creation. To create a flexible adaptive course plug-in, you need to go to the Courses section and select the Add Course option.

Flexible course plug-in was developed from the standard "Topics" course plug-in, as it was consistent with our responsive course model and required fewer programming changes compared to other **formats**: the only element of the course, forum and sections by week.

The structure of the Flexible course is provided by an algorithm that consists of **two phases**: creation and updating. In the first phase, it checks the number of plugins in each section of the course and, if they are not there, adds instances of the TestWidTheory and TestWid plugins to each section using the **completeStructure** function. In the second phase, which occurs when the user adds new sections, our algorithm runs the already mentioned completeStructure function to create the described course structure. Thus, the plug-in mechanism creates a layout for the Flexible course.

6 Description of the design and operation of the TestWidTheory plug-in

As mentioned earlier, the adaptive course plugin consists of two plug-ins: TestWidTheory and TestWid. The plugin for storing educational materials (TestWidTheory) (see Fig. 5) is the prototype of the standard Moodle LMS "Page" plug-in.

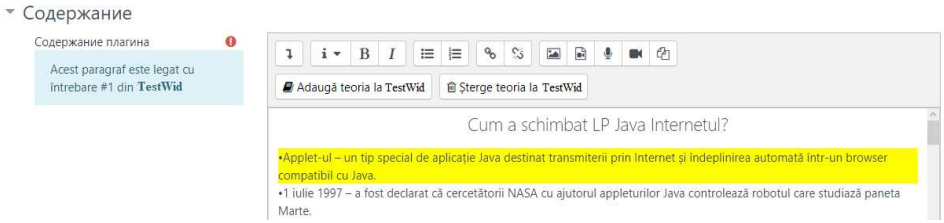


Figure 5. Highlighting a related piece of theory.

This plug-in is a Web-based WYSIWYG editor that is needed to store learning resources (text, video, audio, images) and create links to specific questions from the adaptive test in the current section (see Fig. 2, 3). The work with the plugin for storing educational materials includes the following steps:

1. highlight a fragment of the theory;
2. click on the button "Adaugă teoria la TestWid";
3. in the opened modal window, select the required adaptive test, the question number and click on the "Salvează" button ;
4. Save changes to the plugin.

It should be noted that steps 1 - 3 must be completed 15 times according to the number of questions in the adaptive test. For comfortable work with the plug-in, all related fragments are highlighted when the mouse cursor is hovering.

7 Description of the recovery process and retake

The recovery process is necessary for students with a coefficient $K_a < 0.7$ (grade less than 7), which indicates the incompetence of the students. In this case, after the end of the test, they go to the section with the results. This section can also be accessed through the section "Assessments" (Grades).

In this section, this category of students has the opportunity to study all questions with **partially correct** or **completely incorrect** answers. As it is seen from Fig. 2, each question is accompanied by a link to the relevant theoretical course material. At the same time, the student does not see his previous answers, which makes the recovery process transparent.

After preparation, students take the same adaptive test again. The "TestWid" plugin uses the **load_not_used_questions** function to determine the previously used questions and generate new questions of the corresponding levels. This procedure lasts until either student finally takes the test with satisfactory or bigger grade, or the number of unresolved questions is run out.

8 Testing the developed plug-ins

The study used testing both to create the necessary functions, methods, systems, and to check the quality of their work in general. The final check was carried out on the Moodle Web platform version 3.5.1+ of Alecu Russo Balti State University, and the local one is on version 3.9.1+. Testing was carried out at the following **levels**: unit testing; integrating testing; system testing.

Within each level of testing, the following testing **methods** were used based on manual testing: Installation testing, Usability Testing, Volume Testing, White box, Black box, Grey box, and Graphical user Interface Testing.

The final testing was held on 02.04.2021 at the Alecu Russo

Balti State University, in groups Mathematics and Computer Science (MI21Z) and Computer science (exact sciences) (IS21Z) with overall 26 students, united into one experimental group while studying the course "Programarea orientată pe obiect II (Programarea Java)".

After passing the adaptive test, students passed a questionnaire to assess its work. 18 out of 26 students took part in the survey (69.23%). 55.56% of them agreed that our adaptive testing is better than traditional testing and 44.44% have the opposite opinion.

9 Conclusion

One of the new educational technologies that has shown its undoubted effectiveness is e-learning. After analyzing the principle of operation of the Knewton Alta, Geekie, Smart Sparrow, and NWEA adaptive learning systems, we came to the conclusion to create our adaptive learning system on the base of Moodle Web platform, consisting of three plug-ins: Flexible, TestWid and TestWidTheory.

At the moment, a model of adaptive learning on the Moodle platform has been implemented partially. Only TestWid and TestWidTheory plug-ins functionality has been developed and tested. In the future, our research provides for the full implementation of this model.

Acknowledgments. This article was written within the framework of the research project "20.80009.5007.22 Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] A. Parahonco. *Exploring the capabilities of adaptive learning systems.* In: The Technical Scientific Conference of Undergraduate, Master and PhD Students, vol 1, 2020, pp. 163–166, ISBN 978-9975-45-632-6.

[2] K.A. Vilkova and D.V. Lebedev. *Adaptive learning in higher education: pros and cons.* Modern Education Analytics, vol. 7, no. 37, p.9, ISSN 2500-0608.

[3] *How software that learns as it teaches is upgrading Brazilian education.* [online]. [cited 03.01. 2020]. Available from: https://www.theguardian.com/technology/2016/jan/10/geekieeducational-softwarebrazil-machine-learning.

[4] V. Bocharov and L. Suslova. *Adaptive learning in non-formal education.* In: VI International Scientific and Technical Conference "Modern Information Technologies in Education and Scientific Research" (SITONI-2019), 2019, pp.371–376.

[5] D.A. Bogdanova. *About an adaptive platform for individual training.* In: XI International Scientific and Methodological Conference "New educational technologies in Higher Education", 2014, pp. 202–207.

[6] K. Osadcha, V. Osadchyi, S. Semerikov, H. Chemerys, and A. Chorna. *The Review of the Adaptive Learning Systems for the Formation of Individual Educational Trajectory.* CEUR-WS, 2020, pp. 547–558.

[7] Fundamentals of psychodiagnostics. *Textbook for students of pedagogical universities / under the general editorship of A. G. Shmelev.*, Phoenix Publishing House, 1996, 544 p.

[8] A.V. Solovov, A. Menshikova, and L.Klentak. *Methodological foundations of the design of electronic educational resources: a textbook.* Samar Publishing House, 2013, 180 p., ISBN 978-5-7883-0931-6.

Alexandr Parahonco[1], Mircea Petic[2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Alecu Russo Balti State University
E–mail: alexandr.parahonco@usarb.md

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Alecu Russo Balti State University
E–mail: mircea.petic@math.md

Elearning content processing situations and their solutions

Alexandr Parahonco, Mircea Petic

Abstract

The article discusses processing approaches and their solutions for further usage of content generation in e-courses. It begins with an analysis of web-scraping techniques focused on fetching information from the web network. Then, it discusses their role in modern life and ways of application. Also, the scheme of the model for the dynamic creation of training courses is presented. Finally, the paper discusses content processing situations and their solutions.

Keywords: E-learning, content generation, web scraping, crawler, model.

1. Introduction

The twenty-first century has attended the emergence of groundbreaking information technologies that brought changes in our life. Since the mid-1990s, the Internet gave a start to methods, tools, and gadgets that covered all academic disciplines and business sectors. Soon afterward we witnessed a chain of web 2.0 technologies like E-commerce, which started social media platforms, E-Business, E-Learning, E-government, Cloud Computing, and more other in 2021 [1].

E-learning platforms require the elaboration of high-quality and relevant teaching resources and the constant updating of existing ones. This, in turn, is a complex process consisting of processing a variety of materials, their analysis, synthesis, creative development, and processing of all elements to build a single harmonious structure. Up to now, far too little attention has been paid to dynamic content generation for e-learning courses.

The aim of the article is to analyze web-scraping techniques focused on fetching information and propose a model for the dynamic creation of training courses. Especially it should be noted the discussed content processing situations and their solutions[1].

The article begins with an examination of web-scraping techniques for retrieving data from the Internet. It discusses their role in modern life as well as possible applications. The model's scheme for dynamically creating training courses is then presented. Finally, the paper discusses content processing problems and solutions.

2. State of the art

Udit Sajjanhar was one of the first who wrote an article about extracting information from the Internet in 2008. The author describes educational content mined from university websites in the form of course pages. His system tries to learn the navigation path by observing the user's clicks on as few example searches as possible, and then uses the learned model to automatically find the desired pages using as few redundant pages fetches as possible. Following that, H.W. Hijazi and J.A. Itmazi use keywords divided into two categories as a basis for launching web crawlers: included and excluded from the search query. The authors crawl websites of open educational resources (OER), mainly the Massachusetts Institute of Technology (MIT), which first announced plans to make all of its course materials freely available [2 - 4].

However, the idea of processing web content automatically, technically, came in 1993 with the World Wide Web Wanderer, the first web robot, the sole purpose of which was to gauge the size of the web. Though, it gave birth to the first concept of web-crawling. Soon afterward, the first crawler-based web search engine, JumpStation, was developed. It built a new milestone in web technologies — the prototype of Google, Bing, Yahoo, and other search engines on the web today. In 2004 it acquired a new concept — visual web scraping, provided by BeautifulSoup HTML parser and Web Integration Platform version 6.0

[1] 20.80009.5007.22 Intelligent information systems for solving ill-structured problems, processing knowledge and big data. https://www.math.md/projects/20.80009.5007.22/

[5]. This brought popularity to that technique. Many people, including researchers, started the use of web scraping in different domains.

During the ages, the concept of fetching information from the Internet has evolved into new technology — web scraping (web data extraction). It includes two categories of techniques, such as manual equipment (copy-paste) and automatic data scraping. Manual scraping involves copying and pasting web content, which takes a lot of effort and is highly repetitive in the way it is carried out. Automated scraping techniques shifts from HTML Parsing [6 - 7], DOM Parsing, and XPath to Google Sheets and Text Pattern Matching. Moreover, some semi-structured data query languages, such as XQuery and HTQL, can be used to parse HTML pages and retrieve and transform page content [7].

3. Program model for the dynamic creation of training courses

Our numerous studies guided us to the development of content generation applications. We took as a basis the concept of manual scraping and designed finally the program model given in Fig. 1.

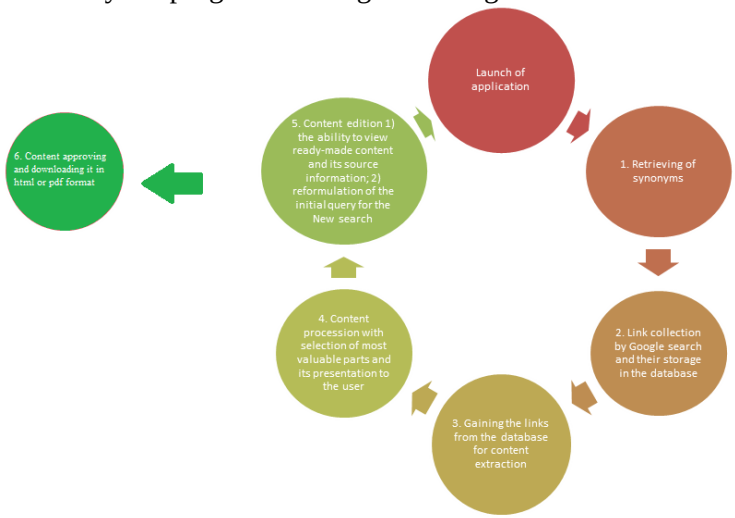


Figure 1. Scheme of the program model for the dynamic creation of training courses

According to our approach, we have 6 steps. In the first step, some web crawlers create networks of synonyms. In the second step, our application uses the original request and/or their selected synonyms for advanced search using Google search. Next, in step three, we gain from Google links for the requests and process them (crawl, select necessary fragments), storing all the information in the database. According to steps 4 and 5, well-merged content should be generated and further exported in HTML or PDF formats in step 6.

4. Resource processing

Resource processing relates to step 3 where the application retrieves links from the database and fetches information from them. This procedure includes text, image, and video extraction along with a selection of meaningful information (due to the search request). A detailed view of this procedure can be demonstrated in Fig. 2.

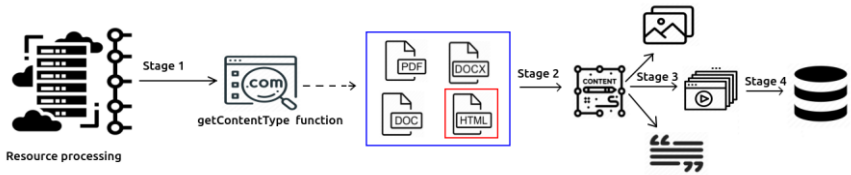


Figure 2. Content selection

The system itself comprises 4 stages for each resource. It begins with the identification of the resource kind. **getContentType** function gets headers of the *url* address and returns its MIME type. Then the necessary modality of handling resources executes. Here a simple web page should be separated from other kinds of sources representing uploaded documents on the Internet. Thus, the content of the web page can be obtained immediately in contrast with resources such as *doc*, *docx*, and *pdf*, which must be first downloaded and then processed.

Regarding web page resources, the retrieved content at stage 2 is used for searching image and video addresses by their tags *img* and *video at stage 3*. Text extraction (meaningful parts) is undertaken by **findInText** function. All gained information then is stored in the database.

Documents such as *doc*, *docx,* and *pdf* have another sequence of actions. After the download, they get processed. *Doc* format contains binary data and the application cannot obtain all the content except for the text. Hence, **findInText** function is applied at the next stage to find the searched information and save it. Both *docx* and *pdf* formats are more flexible in this respect as they allow us to fetch images and text. The procedure with text does not differ from *doc* format. However, throughout the image manipulation, the application creates some folder structures on the server and stores them there. Next, each image undergoes the procedure of optical character recognition (OCR). This is important as often information is scanned and saved into *pdf* or *docx* formats (books, magazines). Thus, OCR provides some text to find with **findInText** function.

At the end of resource processing, results are presented to the user.

5. Conclusion

The huge amount of information on the Internet is the mark of the 21st century. There is similar information, but it is outdated or fake. This necessitates the involvement of information technologies and approaches in order to select the most valuable pieces of it. It is especially important in the education area, where qualitative and up-to-date content plays an important role in the specialist formation.

The research proposes to build a focused web crawler for intelligent data extraction from web sources. We have adopted this idea and proposed an application model for the dynamic creation of training courses. Currently, we have focused on content processing in different file types and have elaborated on a common algorithm for its processing. Our next step is to implement Data mining and Text mining approaches to give the generated content more logic.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

References

[1] M. Yamin. *Information technologies of 21st century and their impact on the society*. In: “International Journal of Information Technology”,

vol. 11, no. 16-8 (2019), pp. 759-766, DOI: 10.1007/s41870-019-00355-1.

[2] U. Sajjanhar. *Focused Web Crawling for E-Learning Content*. Master's thesis. Indian Institute of Technology Kharagpur, Master of Technology In Computer Science and Engineering, 2008.

[3] H.W. Hijazi and J.A. Itmazi. *Smart Crawler Based e-Learning*. Hammamet-Tunisia, vol. 11, 2013, pp. 209 – 2016.

[4] Y. Biletskiy, M. Wojcenovic, and H. Baghi. *Focused Crawling for Downloading Learning Objects – An Architectural Perspective*. Interdisciplinary Journal of E-Learning and Learning Objects, vol. 5, no. 1, 2009, pp. 169 – 180.

[5] D. Kremer. *Brief History of Web Scraping*, 2021, https://webscraper.io/blog/brief-history-of-web-scraping.

[6] Bo. Zhao, *Web Scraping. Springer International Publishing*, vol. 5, (2017), 632 p, DOI: 10.1007/978-3-319-32001-4_483-1.

[7] S. Ruihua. *Joint Optimization of Wrapper Generation and Template Detection*. KDD07: The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Minin, August, (2007), pp. 894 – 902, DOI: 10.1007/978-1-4419-7735-9_4.

[8] *Introduction to Web Scraping*, 2019, https://www.geeksforgeeks.org/introduction-to-web-scraping/.

Alexandr Parahonco[1], Mircea Petic[2]

[1]Affiliation/Institution: junior researcher / Vladimir Andrunachievici Institute of Mathematics and Computer Science, MD-3100, Moldova

software engineer / Department of Information Technologies, Alecu Russo State University of Balti, MD-3100, Moldova
E-mail: alexandr.parahonco@usarb.md

[2]Affiliation/Institution: leading researcher / Vladimir Andrunachievici Institute of Mathematics and Computer Science, MD-3100, Moldova

PhD, associate professor / Alecu Russo State University of Balti, MD-3100, Moldova
E-mail: mircea.petic@math.md

The model of Web crawler for expansion the scope of initial search

Alexandr Parahonco, Mircea Petic, Corina Negara

Abstract

The paper discusses present obstacles of content formation and its maintenance while teaching process and offers an alternative modern way for this purpose. The article proposes the program model of the new application for creating dynamic content of training courses and briefly reviews the process of dynamic generation of teaching content and its integration into the Moodle learning platform. It discusses the construction of custom web-crawler and its functions within the program model.

Keywords: content generation, program model, Moodle, web-crawler, web-scraping, e-learning.

1. Introduction

One of the new educational technologies that has shown its undoubted effectiveness is electronic education, or in the original transcription - e-Learning. In developed countries, e-learning covers all levels of education and is widely used not only in universities, but also in high school and in the organization of corporate (postgraduate) education [1, pp. 1367-1368; 2, pp. 3–4]. Almost all universities and most American schools have implemented E-Learning. Increasing penetration of internet in many regions across the globe is a major factor driving the market growth [3, p. 8].

Such platforms require the elaboration of high-quality and relevant teaching resources, the constant updating of existing ones. This, in turn, is a complex process consisting of processing a variety of materials, their analysis, synthesis, creative development and processing of all elements to

build a single harmonious structure. Up to now, far too little attention has been paid to dynamic content generation for e-learning courses.

The aim of this work is to describe the developing application model for creating dynamic content of online training courses in Romanian, English and Russian languages, and its first part - the development of a search robot ("crawler") for expansion the scope of initial search.

An introduction part covers short description of the research, its structure and aims; literature review provides the problem actuality and solutions found by researches. Then we speak about our version of the problem sort out – new system of content generation for e-courses. Further we bring empirical model of our crawler along with experiment results and their analysis.

2. Literature review

The use of internet technology to deliver educational content is the latest trend in training and education development industry. A Learning Management System (or LMS) is often used to manage user learning processes. However, the majority of the current online based learning systems has two serious drawbacks: 1) non availability of ready content what leads to a dead end an instructor who begins to make up a course without the material to start up and 2) the rapid changes in the educational content, the vast amount of published papers, and the ever increasing training tutorials that necessitate the dynamic update of the existed courses in e-learning system.

The literature review led us to the solution proposed in [4, 5, 6]. These researches suggest one to use focused web crawler as a way to gather and process information from the internet. However, questions have been raised about the crawler architecture.

In [4] the researchers take as a basis for launching web crawler key words divided in two categories: included and excluded from the search query. The authors solve the first and the second problems by crawling the websites of open educational resources (OER) and mainly Massachusetts Institute of Technology (MIT) which first announced to make all of its course materials freely available. Nowadays this list is extended.

The web crawler is in charge of traversing the academic websites of open educational resources; retrieving the content and indexing it

according to the keywords. The administrator specifies both open academic web sites and keywords based on the e-course ID. The next step consists in parsing stored websites and extracting the content from the keywords related pages [4].

What is less clear is the nature of content selection approach. There is nothing said whether the downloaded e-courses are taken as a whole part or undergo some selection procedures of content. Thus, we can only assume that content is received entirely.

In [5] the authors delve into pedagogy and work out the upper mentioned problems by learning object metadata (LOM). Learning objects (LO) are learning resources named in conformance with the object-oriented paradigm that may be used, re-used or referenced during technology-supported learning. Thus, LOM is a data model, usually encoded in XML, used to describe a learning object and similar digital resources.

The scientists rely on digital libraries and learning object repositories, such as NEEDS and SMETE to search and download, at first, LOM storing in the local repository; then accessing its file with actual learning object content by reference (URL) within LOM. Such framework decreases the time for search and delivery of learning objects to learners.

The Focused Web Crawler system consists of two main applications – ID Web Crawler and LOM Downloader. ID Web Crawler takes responsibility for making a list of URLs containing LOM instances. Thereafter, the second application – LOM Downloader parses the pages, gets LOM instances and retains URL of the file. Whereby it obtains learning materials from different resources and composes one course.

Like in the previous article, this crawler cannot filter LO and gets only the necessary part of it and is compelled to use whole resource that is not the optimal solution [5].

Another article [6] throws light on educational content mined from University websites in the form of course pages. Researches claim that content can be mined from the following sources:

(a) Websites and open source course material like MIT Open Courseware, NPTEL India.
(b) Course websites of large international universities from the USA.
(c) Discussion Forums - Google Groups, Yahoo Answers

(d) Websites for animations/videos - Youtube, Google Video and metacafe
(e) Websites for general content - Wikipedia, Mathworld.

Their system tries to learn the navigation path by observing the user's clicks on as few examples searches as possible and then uses the learnt model to automatically find the desired pages using as few redundant page fetches as possible [6].

Considering the proposed crawler architecture, we decided to use similar web crawler model as it was suggested in [4, 5, 6]. The difference will consist of both expansion - keywords area and search zone. As our application should serve for multiple educational domains, neither open source courses nor digital libraries are available adequately in the Internet, on the one hand, and do not amount to considerable volume in Romanian and Russian languages, on the other. Therefore, we decided to use essential search zone – Google search engine. In order to simplify the search process for the user and expand keywords area we lean towards the idea of using dictionaries of synonyms, which may create various search requests and bring qualitative content. Consequently, we need two crawlers: one for parsing dictionaries of synonyms and another for working with Google engine. This article describes focused crawler of the first type.

The development of web crawler led us to web scraping technique. As a matter of fact, web scraping – or web crawling, were historically associated with well-known search engines like Google or Bing. These search engines crawl sites and index the web. Because these search engines built trust and brought back traffic and visibility to the sites they crawled, their bots created a favorable view towards web scraping. It is all about how you web scrape and what you do with the data you acquire. Nowadays there is no need to be an expert programmer to scrape web data. There are software solutions that render alike services: Impoprt.io, Octoparse, ScrapeSimple, ParseHub etc. Nonetheless, it should be mentioning that in case if it is compulsory to further process data (cleaning, deduplication, etc.) a web scraping software cannot really help [6, p. 2; 7, p. 2].

There are several techniques for using web scraping. Two categories can be distinguished among them: manual equipment (copy-paste), and automatic.

Manual scraping involves copying and pasting web content, which takes a lot of effort and is highly repetitive in the way it is carried out.

Automated scraping techniques shift from HTML Parsing [8, pp. 1 – 2, 9, p. 894], DOM Parsing, XPath to Google Sheets and Text Pattern Matching. Moreover, some semi-structured data query languages, such as XQuery and the HTQL, can be used to parse HTML pages and to retrieve and transform page content [10, p. 10].

All these techniques are good but we want to perform an automated scraping in most advanced and customized way using PHP server language. Hence, Google Sheets and Text Pattern Matching should be foreclosed from the list of candidates.

From the remained techniques most of all for our application fits "HTML Parsing" as we are planning to parse web pages by elements of the markup and CSS languages according to the template. Moreover, it is worth noting that even the automated web scraping process requires manual configuration to create a data extraction template [11, p. 365; 12, p. 136].

This article reports about elaboration of developing application model for creating dynamic content of online training courses in Romanian, English and Russian languages, starting from development of a crawler for expansion the scope of initial content search query.

3. Analysis of the proposed structure

Our literature review has led us to the idea of expansion search area of dynamic content generation. We started from the model of manual scraping and came finally to the program model represented in Figure 1.

As it can be seen at Phase 1 the operating principle of the developed model is to use synonymous connections to search for dictionaries that are similar in meaning with the help of a crawler, and use them at Phase 2 and 3 for advanced search using the Google search service. Thus, we obtain the behavior model of a user performing manual scraping.

According to phase 4 and 5, well merged content should be generated and further exported in Moodle. The content may be imported in Moodle

via Page and File standard plugins at Phase 6. Our application will allow to edit content (font, placement, color) and to download the files in html or pdf formats.

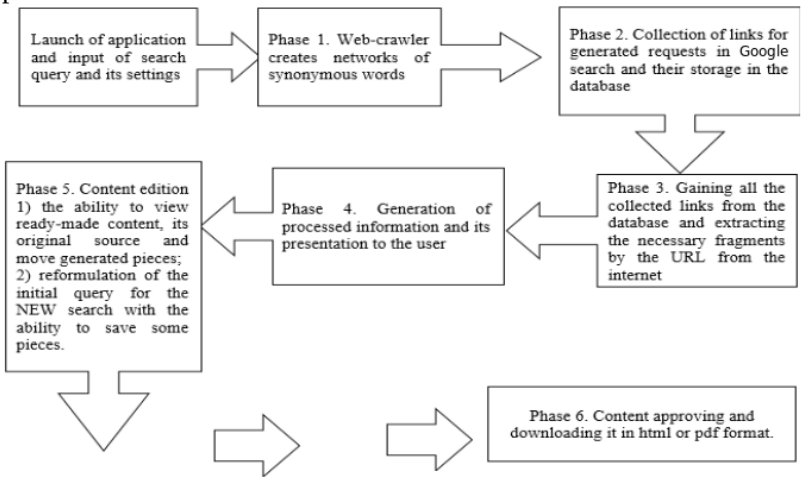


Figure 1. Scheme of the program model for the dynamic creation of training courses

4. Model of web crawler.

According to Phase 1 the first significant element in this model is the creation of a web crawler based on the server-side programming language PHP and Goutte screen scraping and web crawling library [13, p. 4]. This mechanism should create a semantically-related network of terms for a larger set of search query options.

Firstly, let's look at the main page that launches the application. As it can be seen from the picture (Figure 2) we have a "text input" for the search query, one "select input" for language type of dictionary and a start button. It is our first sample of the page and further it surely will be enriched with different options.

Secondly, we designed a table "dictionaries" that stores information about dictionaries with synonyms. When user launched the application, the dictionaries are selected by the language parameter type ("ru", "ro", "en").

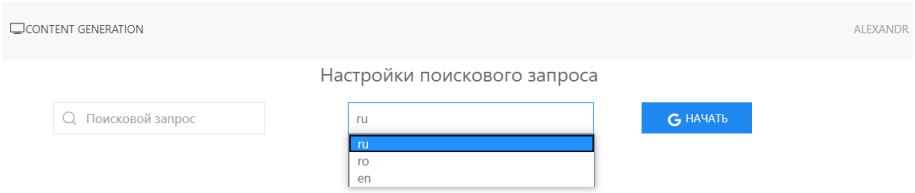


Figure 2. Main page of the application

The next step addresses crawler logic and elaboration of synonymous network. It can be divided into four phases:

1. extraction from query string prepositions and articles that do not make sense in finding synonyms (filtering);
2. reception of records from data table “dictionaries”;
3. passing through the loop of words from the filtered search string and loop of dictionaries (Figure 3);
    a. connection of crawling template;
    b. crawling the dictionary page by words iteration;
    c. extraction of synonyms;
    d. removal of non-unique words;
4. design of the data set structure.

```
foreach ($search_query_prepared as $index => $word) {

    foreach ($dictionaries as $key => $dictionary) {
        $param = ($dictionary->param_type == 'question')? '?' : '/';
        $url = $dictionary->url . $param .$dictionary->link_param .'=' . $
            word;
        $crawler = $client->request('GET', $url);
        $words = '';
        require(__DIR__. '/../Templates/' . $dictionary->template);
        if(!empty($words)){
            $extended_search_query[] = [
                'index' => $index,
                'synonyms' => implode(',',array_unique(explode(',', $words)
                    )),
                'original' => $word
            ];
        }
    }
    //check if there are synonyms
    if(strlen($words) > 0){
        //there are some synonyms
        Log::info('Found synonyms', ['synonyms' => $extended_search_query])
            ;
        Cache::put('synonyms', $extended_search_query);
    }
}
```

Figure 3. Crawler logic and design of data set

The data set structure represents an array of synonyms. Each element represents an array with index (position within search query), comma

separated synonyms and original word. This data set is further cached for future processing. We find it rational to store synonymous network in cache rather than in Data Base based on the resource economy, the speed of access and aim.

Generally, Data Base requires storing data for long time. Cache services were designed to store data for short time, although permanent storage is also available. In our case synonyms do not represent any value for retaining as their application will take long time and special algorithms. Online dictionaries already have words network for prompt search of synonyms. This is the main point to use them instead of Data Base. Furthermore, our aim is to build an application that does not consume huge resources for maintenance [14, p. 3 - 4].

As part of initial development, we have elaborated user interface for displaying synonyms and selected one dictionary – "Wordsmyth The Premier Educational Dictionary-Thesaurus" [15]. Our examination of the site and its structure persuaded us to take similar words category as a source for data set structure if synonyms category is absent. The application page provides accessible for understanding way to display original word from search query and its synonyms. It also shows the query itself. From it we can see that words as "of" and "in", being pronounced, were omitted from crawling approach, entering the list of exceptions.

Our experiments demonstrate that variations of our query expand the scope of the initial search. For example, the search for "abolition of serfdom in Russia" provides the following results (Figure 4) and Figure 5 provides the results for manual search for the word "abolition".

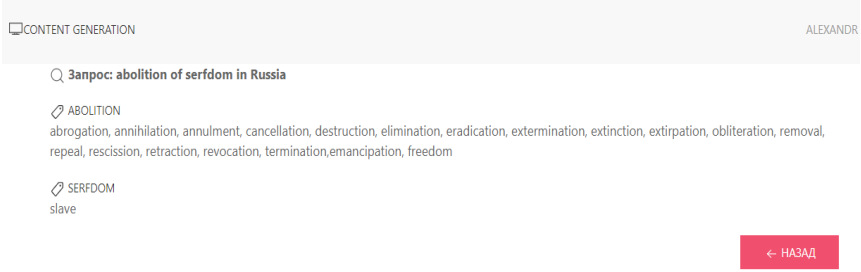


Figure 4. Web crawler and web scrapping results

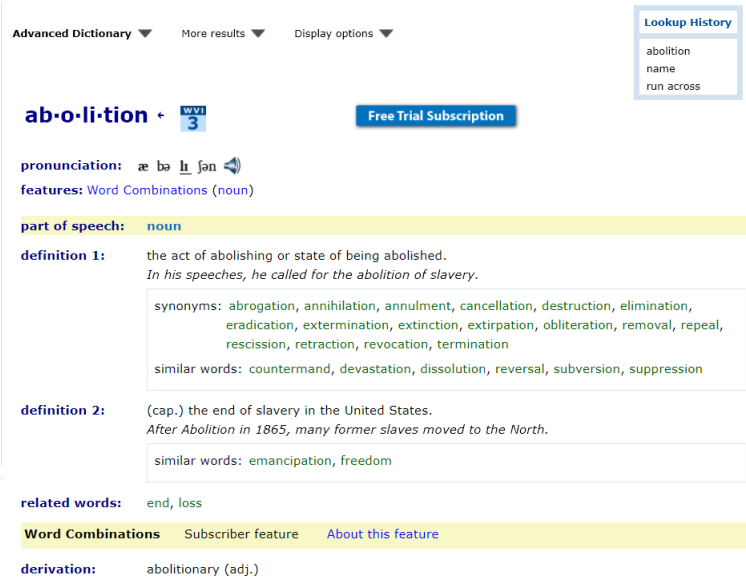


Figure 5. Wordsmyth page result for manual search for the word "abolition"

As it can be seen from the current example and others (Table 1), variations of our query are obtained that expand the scope of the initial search. However, it should be emphasized that there are also "irrelevant" combinations of inquiries (annulment, obliteration, etc.) which are non-commonly used in digital sources in the context of the search query and probably will lead to false searches. However, this effect is partly leveled at the stage of searching by Google, having at hand full query and can be reduced by preliminary selection of fragments with the desired text.

Table 1. Application examples of synonym network.

Initial query	Synonyms
design patterns in software engineering	DESIGN engineer, map out, plan, conceive, dream up, formulate, invent, earmark, intend, target, contrive, devise, blueprint, draft, layout, program, scheme, strategy, drawing, picture, pattern, art, drawing, fashion, graphic arts, graphics, painting, aim, goal, intention, objective, purport, purpose, intentions, plot

	PATTERN design, figure, motif, configuration, shape, structure, archetype, ideal, model, paradigm, prototype, standard, model, style, form, kind, rhythm, tempo, type, variety, fashion, figure
Introduction to linguistics	INTRODUCTION establishment, inauguration, initiation, insertion, institution, interposition, presentation, presentation, foreword, overture, preamble, preface, prelude, proem, prologue, innovation, overview, survey

5. Conclusions

In the course of our research we found out that today's e-learning environment suffers from the cold start problem – lack of readymade content for courses. Experts suggest building the focused web crawler for intelligent data extraction from web sources. We have adopted this idea in our application model and enhanced it by widening keywords area by composing semantically-related network of word and search zone by Google engine.

Further literature review brought us to the web-scraping technique that speaks about ways of content extraction and we selected "HTML Parsing" for our focused web crawler.

The empirical path starts from the elaborated program model of our application. It includes six phases. At Phase 1 we use synonymous connections to search for dictionaries that are similar in meaning with the help of a crawler, and use them at Phase 2 and 3 for advanced search using the Google engine. Thus, we obtain the behavior model of a user performing manual scraping. According to phases 4–6 the well merged content should be generated and further exported in Moodle via Page and File standard plugins.

As expected, we succeed at expansion the scope of the initial search query and produced semantically-related network of terms that is stored in the cache rather than in database. Insofar as cache is fast and resource-saving solution.

During the tests we observed that by crawling the dictionaries we gained many repeated words that we had to remove keeping only unique synonyms. The omitted words in the examples (Russia, software,

engineering, linguistics) do not have synonyms in the dictionary used. However, this does not matter as during the phase 2 the initial words will be included in search query at the beginning and then step by step will be substituted by their synonyms if they have.

Finally, our research is passing to the next stage – implementation of the Phase 2 and 3. We should now use our synonyms and collect links in database for their further crawling and extraction of "valuable" content.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] D. Benta, G. Bologa, and S. Dzitac. *University Level Learning and Teaching via E-Learning Platforms*. Procedia Computer Science, no 55 (2015), pp. 1366–1373.

[2] J. Valverde-Berrocoso, M. Garrido-Arroyo, and C. Burgos-Videla. *Trends in Educational Research about e-Learning: A Systematic Literature Review (2009–2018).* Sustainability, volume 12, no 5153 (2020), pp. 1–23.

[3] L. Wood. *World Online Education Market, Forecast to 2025 by Type, Technology, Vendor, End-user and Region*. Global Online Education Market - Forecasts From 2020 To 2025, no 4 (2020), 124 p.

[4] H.W. Hijazi and J.A. Itmazi. *Smart Crawler Based e-Learning*. Hammamet-Tunisia, volume 11 (2013), pp. 209–2016.

[5] Y. Biletskiy, M. Wojcenovic, and H. Baghi. Focused Crawling for Downloading Learning Objects – An Architectural Perspective. *Interdisciplinary Journal of E-Learning and Learning Objects*, volume 5, no 1 (2009), pp. 169–180.

[6] U. Sajjanhar. *Focused Web Crawling for E-Learning Content: master's thesis. Indian Institute of Technology Kharagpur*, Master of Technology In Computer Science and Engineering, 2008. https://pdfs.semanticscholar.org/2da0/a0658ce31e2143cec9a050aebc3337cb188.pdf.

[7] A. Toth. *Is Web Scraping Legal? 6 Misunderstandings About Web Scraping*, *Scraping Authority, no 11*, 2017. https://www.import.io/post/6-misunderstandings-about-web-scraping/

[8] Bo. Zhao. *Web Scraping*. Springer International Publishing, Volume 5, (2017), 632 p.

[9] S. Ruihua. *Joint Optimization of Wrapper Generation and Template Detection*. KDD07: The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Minin, August, (2007), pp. 894–902.

[10] *Introduction to Web Scraping*, 2019 https://www.geeksforgeeks.org/introduction-to-web-scraping/

[11] Anand V. Saurkar, Kedar G. Pathare, Shweta A. Gode. *An Overview of Web Scraping Techniques And Tools*. IJFRCSCE, volume 4, no 4 (2018), pp. 363-367.

[12] S. Sirisuriya. *A Comparative Study on Web Scraping*. Proceedings of 8th International Research Conference, KDU, no 11 (2015), pp. 135 – 140.

[13] V. Bambuch. *Platform for Cryptocurrency Address Collection*. Excel 2020 Sbornik, no 6 (2020), pp. 3–10.

[14] I. Haber. *15 Reasons to use Redis as an Application Cache. White Papers, 2016.* https://redislabs.com/wp-content/uploads/2016/03/15-Reasons-Caching-is-best-with-Redis-RedisLabs-1.pdf

[15] *Wordsmyth The Premier Educational Dictionary-Thesaurus*, 2020 https://www.wordsmyth.net/

Alexandr Parahonco[1], PeticMircea[2,1], Corina Negara[1]

[1] Alecu Russo Balti State University, Republic of Moldova
E-mail: alexandr.parahonco@usarb.md, corina.negara@usarb.md

[2] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Republic of Moldova
E-mail: mircea.petic@math.md

Part 5

Systemic concept of the heterogeneous multi-cloud platform and methods of realizing the execution environment of imaging information processing applications

Upgrading Cloud Infrastructure for Research Activities Support

Nichita Degteariov, Petru Bogatencov, Nicolai Iliuha, Grigore Secrieru

Abstract

In the paper, approaches for upgrading the heterogeneous distributed computing infrastructure that integrates various types of computing resources are described. It is shown the necessity of development of computer infrastructures and technologies, which is focused on creating conditions for solving complex problems with high demands of computing resources. Analysis and trends of development of tools for automation of complex cloud infrastructures configuration and administration are presented. Problems that restricted scalability of the existing Cloud infrastructure are identified and solutions to overcome existing limitations by application of new tools for cloud infrastructure configuration and administration are suggested.

Keywords: cloud computing, information technologies, e-infrastructure & services, deployment tools.

1. Introduction

In the past years, development of distributed and high-performance computing (HPC) technologies for solving complex tasks with specific demands of computing resources, creating abilities to store and access increasing amounts of research data are actively developing, including in Moldova [1]. New European Open Science Cloud Initiative (EOSC), aimed at the accumulation of various scientific information in cloud for organization of open access, has a further significant impact on the intensification of the use of distributed computing resources. The initiative is oriented at creation of open research data repositories to support open

science and development of technologies for the accumulation and use of FAIR (Findable, Accessible, Interoperable and Reusable) data based on wide utilization of cloud computing resources. New areas of works in this direction focused on deployment of new types of cloud infrastructure that will integrate Grid and HPC computing resources and gain benefit to end users from uniting computational resources of multiprocessor clusters with effective application platforms, users' interfaces and infrastructure management tools offered by Cloud infrastructure.

2. Overview of Cloud Infrastructure development for research activities support at the Vladimir Andrunachievici Institute of Mathematics and Computer Science (VA IMCS)

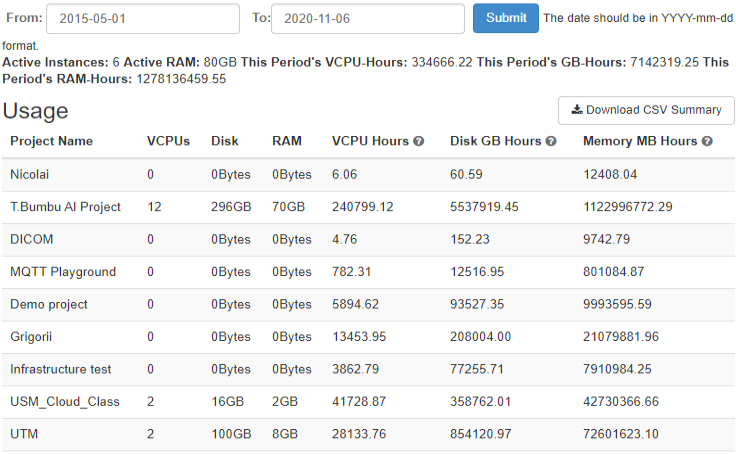
Work on the implementation of cloud infrastructure at the Institute of Mathematics began in 2014-2015, as a result of the participation of VA IMCS and partner organization RENAM engineers in the regional project Experimental Deployment of an Integrated Grid and Cloud Enabled Environment in BSEC Countries on the basis of gEclipse (BSEC gEclipseGrid) supported by Black Sea Economic Cooperation Programme (http://www.blacksea-Cloud.net) [2].

The experience and results accumulated in the course of this project, in 2015, resulted in the Cloud infrastructure available for evaluation and application testing, based on OpenStack version 13 Mitaka (release 2016), installed and accessible at https://cloud.renam.md/. It was deployed using two computing nodes of the multiprocessor cluster of the Vladimir Andrunachievici Institute of Mathematics and Computer Science, which at that time were taken out of mainstream operation in this computing facility. The total amount of resources on these two servers was quite modest even by those standards - only 16 CPU cores, 32 Gb RAM, 750 Gb HDD, 1 Gbit/s network. Despite this, this infrastructure was widely used, both for evaluation purposes, and in several short-term projects and tasks in which it was necessary to deploy quickly a small virtual infrastructure to test various scenarios and products.

In 2018, the infrastructure was supplemented with a modern high-performance server with 32 CPU threads, 128 Gb RAM, 3 Tb storage, 1Gb network. This upgrade has opened a new stage in the development of Cloud computing facility at VA IMCS – new resources for several

institutional projects in the field of Machine Learning and Neural Language Processing were provided there.

In 2019, the Cloud infrastructure of the Institute of Mathematics was used for new service – support of on-line lectures organization for the State University and the Technical University of Moldova. The experience gained over the years and user feedback made it clear about the need for further development of Cloud infrastructure in VA IMCS (see Figure 1), but a further increase in resources was no longer an option - the technologies used in the existing Mitaka release no longer met modern security standards, and the manual installation process used to implement OpenStack created insurmountable difficulties for further upgrade and system administration.


From: 2015-05-01 To: 2020-11-06 Submit The date should be in YYYY-mm-dd format.

Active Instances: 6 **Active RAM:** 80GB **This Period's VCPU-Hours:** 334666.22 **This Period's GB-Hours:** 7142319.25 **This Period's RAM-Hours:** 1278136459.55

Usage — Download CSV Summary

Project Name	VCPUs	Disk	RAM	VCPU Hours	Disk GB Hours	Memory MB Hours
Nicolai	0	0Bytes	0Bytes	6.06	60.59	12408.04
T.Bumbu AI Project	12	296GB	70GB	240799.12	5537919.45	1122996772.29
DICOM	0	0Bytes	0Bytes	4.76	152.23	9742.79
MQTT Playground	0	0Bytes	0Bytes	782.31	12516.95	801084.87
Demo project	0	0Bytes	0Bytes	5894.62	93527.35	9993595.59
Grigorii	0	0Bytes	0Bytes	13453.95	208004.00	21079881.96
Infrastructure test	0	0Bytes	0Bytes	3862.79	77255.71	7910984.25
USM_Cloud_Class	2	16GB	2GB	41728.87	358762.01	42730366.66
UTM	2	100GB	8GB	28133.76	854120.97	72601623.10

Figure 1. Cloud in VA IMCS 2015-2020 resource usage

3. Identified problems in the existing Cloud infrastructure and suggested solution

The fact is that the OpenStack Cloud infrastructure is a very flexible and, as a result, a very complex product, consisting of hundreds of open-source "bricks" components combined into a single system, which have many dependent on each other services and components, with their own settings, which must be written into the configuration files before executing the

commands to "dock" the components into a single working environment. Mistakes when installing such a complex system are inevitable and sometimes irreversible and necessitate reinstalling all components from scratch. This approach makes further administration and scaling of the system absolutely impractical and complicated process.

Installing OpenStack manually is great for getting familiar with the internals of a system and understanding how its components interact. However, when installing the system on large infrastructures, where clustering of more than 3-5 servers is used, manual installation is already absolutely inappropriate and unpromising in terms of further scaling.

To overcome these limitations, increase the reliability and the possibility of further upgrading and scaling the Cloud system, a completely different approach was required.

To solve this problem, it was decided to apply a modern approach to the administration of cloud systems - using Deployment Tools, which allow creating scripts to automate the installation of the system. It is difficult to imagine a modern IT project without such solutions. There are already known about a dozen of such kind of tools to automate configuration processes, the main ones and mostly used being Ansimble, Puppet, Chef, Juju. To deploy a new improved Cloud system on the basis of modern equipment, we have chosen a combination of opensource tools such as MAAS (Metal-As-a-Service) and Juju Charms.

MAAS is designed to deploy quickly and easily Ubuntu configurations across multiple servers using techniques used in cloud platforms. But unlike cloud platforms, resource allocation on such kind of cluster occurs at the level of physical servers, not virtual environments. At the heart of MAAS there is the simple idea of Preboot eXecution Environment (PXE) booting and a tool for deploying and maintaining Juju environments, which turns the installation and configuration process into an extremely simple task, performed using two or three commands [3]. It would take too long to manually configure the OS and services on each server node in the cluster, whereas tools like MAAS can deploy an entire cluster in just a few minutes.

Juju is an orchestrator that can be used to declaratively describe the infrastructure configuration of an application: which applications are running, on which machines, in how many copies, and how they are

linked to other services. The custom code for configuring individual virtual machines with Juju is called Charm [4].

Having the installation of automation system, we get not only a gain in man-hours for deploying a ready-made production-ready infrastructure, but also flexibility in its administration and ease of scaling. For example, having a Compute Node setup script written, you can start provisioning any number of new nodes with a single command, using an existing debugged configuration. This allows to minimize the occurrence of errors when commissioning new system components and to carry out maintenance with minimal delays or even without downtime.

Creation of new infrastructure will allow us to eliminate many of the limiting factors of the infrastructure currently operating in our research Cloud. The main advantages of the updated system are: more computing resources, block-storage for creating a backup, a new network model that allows users independently to create self-service local networks with local addresses and use the mapping of floating IP address. This, in turn, will increase security and significantly reduce the use of public IP addresses – for example, now only 16 real IPs are available and used in the current system.

4. Future plans of research Cloud infrastructure development

Introduction of new technologies, ideas and planned improvement of the computer and network infrastructure in the VA IMCS and RENAM data centers gave impetus to the creation and transition to a new, more modern, productive and large-scale Cloud platform.

The new Cloud platform is designed to eliminate bottlenecks in the current system, provide users with more resources using modern high-performance servers, more bandwidth by migrating infrastructure to 10 Gbps connectivity, and being more reliable, flexible and resilient by using automated deployment tools.

This work began at the end of 2020, with the transition of the VA IMCS and RENAM infrastructures to new high-performance servers and new 10Gbps network equipment. Completion of the work is planned in 2021, and then the system will become available for testing and after a successful evaluation of results the obtained VA IMCS resources will also be integrated in the upgraded Cloud infrastructure.

5. Conclusion

Today it is already impossible to imagine life without Cloud systems. They penetrate all areas of our lives and continue to gain popularity, and the current state of affairs with the coronavirus pandemic will continue to increase the demand for the provision of more and more virtual resources and services in the Cloud. They will be an excellent aid for supporting scientific and educational activities.

Acknowledgments. This work was supported by research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" and EU H2020 project "National Initiatives for Open Science in Europe – NI4OS Europe" (Grant Contract 857645).

References

[1] P. Bogatencov, G. Secrieru, N. Degteariov, and N. Iliuha. *Scientific computing infrastructure and services in Moldova*. Springer Link, Journal Physics of Particles and Nuclei Letters LNCS, vol. 13, Issue 5 (2016), pp. 685-688, DOI: 10.1134/S1547477116050125.

[2] H. Astsatryan, A. Hayrapetyan, W. Narsesian, P. Bogatencov, N. Iliuha, R. Kvatadze, N. Gaamtsemlidze, F. Florian, G. Neagu, and A. Stanciu. *Deployment of a Federated Cloud Infrastructure in the Black Sea Region, Computer Science and Information Technologies*. Proceedings of the CSIT Conference, Sep. 23-27 — Erevan, Armenia (2013), pp. 283–285.A.A. Waksman. *Permutation Network.* Journal of the ACM, vol. 15, no 1 (1968), pp. 159-163.

[3] *MAAS documentation* — https://maas.io/docs.

[4] *Juju documentation* — https://juju.is/docs.

Nichita Degteariov[1], Petru Bogatencov[1], Nicolai Iliuha[1], Grigore Secrieru[1]

[1] Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: nichita.degteariov@math.md

Incident Handling and Personal Data Protection in Medical Images systems

Alexandr Golubev, Petru Bogatencov, Grigore Secrieru, Ecaterina Matenco

Abstract

Modern e-Health systems require innovative solutions for data protection and overall information system security on the one hand and mechanisms for data share, taking into account personal data protection laws and regulations on the other hand. This article will show specific problems for data protection for Medical images collected in DICOM format and propose solutions for security incident reporting and tracking. Main issues that should be taken into account and addressed is the fact that not only metadata could be treated as personal data, but so is for any data that can be used to identify a person, because the latest software solutions could restore human face based on set of X-ray films that can be used to identify a person. Those tasks for data protection can be solved by using innovative ticketing and monitoring systems that are described in this article.

Keywords: Information Security, Ticketing system, Medical Images, DICOM, Personal Data, CERT.

1. Introduction

E-health systems working with medical images play important role for modern hospital information systems (HIS) all over the world. In Moldova, there exist three popular HIS realizations and many custom small medical informational systems installed both in private and public hospitals and diagnostics centers. Most of those systems have integration

with various types of medical images collections that are connected to the patient medical record.

Personal patient data is the most sensitive and important information that should be secured when it is stored in these systems. All patient personal data are protected by NCPDP (The National Center for Personal Data Protection of the Republic of Moldova) regulation that is based on national legislation. That means that patient data could not be shared or transferred without special agreement with the person or other owners of his data. That's why patient should sign a special agreement when he is visiting hospital, where he/she agrees that the hospital will use the data inside the institution to treat the patient, but will not share any patient information outside the institution.

That creates patient's data security problems in most cases when any medical information is used for external consultations, when analysis or medical images should be transferred to another medical institution/doctor or even when the patient wants to take medical images away. That obviously creates preconditions for necessity to open public hospital portal where patient should confirm that he allows sharing his information online without issuing any additional paper-based agreements.

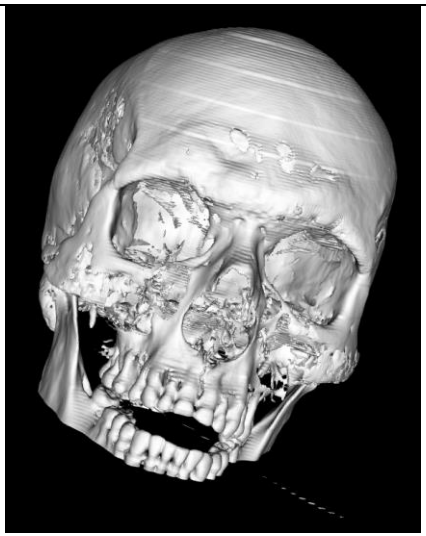
This article is targeted mainly on medical images security, so the attention will be focused on technical aspects of medical image sets processing for secure transfer and procedures required in case if any data security incident occurs.

2. Specific issues for DICOM images security and data protection

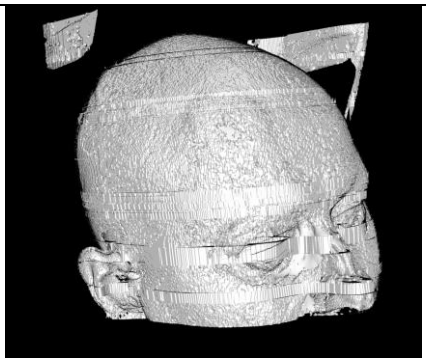
The common standard for medical images storing and transferring is DICOM format [1]. This format includes metatags that are built in the file using XML format and image itself in raw high-resolution format.

Based on conclusions made by authors of this research, it is highly recommended to remove metadata information from medical images after image is processed and stored in HIS or in Laboratory information system (LIS), as far as this data is not important for medical purpose data processing, but contains personal patient information that is not possible to protect in case if file will be stolen or lost.

By removing the metadata, it is possible to reduce available volume of personal data, but unfortunately, it will not exclude possibility of patient identification. In the Figure 1 below it is shown the "Bones" layer that is built based on simple data transform and more sophisticated "Soft Tissues" that could be used for face reconstruction. That is possible because each raw data pixel contains information about the "material" that it represents in the human's body:



Bones



Soft Tissues

Figure 1. "Bones" layer simple data transformation for "Soft Tissues" restoring

The other specific issue is that data share and transfer is possible, allowed and in many cases even required for modern e-health system, so it is the real scenario when a patient will require his data to continue the treatment in another medical institution [2]. That requires realization of a mechanism and specified procedures for hospital and patient that will allow request data transfer and trusted approval mechanism that will ensure correct identification of the requester and verified approval procedure. One of the best mechanisms could be digital signature based

on personal digital certificate like “MSign”, that is linked to the person and have confirmed juridical value.

The last but not the least is data anonymization that will make possible share data for research and educational purposes. As it was described in the beginning of this section, any image can be considered as personal data, so patient should approve using his data before it could be widely used by researchers and for educational purposes [3]. This will also require creation of specific approval mechanism that will use trusted and protected tools and procedures [4].

3. Incident Handling in modern Ticketing systems

The issues described in the second section could and should be solved by specific algorithms and tools built in Medical Information System (MIS), but also should be monitored and analyzed both automatically by monitoring tools and by security officers. This will make necessary to collect and handle all security incidents. By term “incident” we mean any issue or request that could not be solved automatically. This could be a request to close access, or alert from monitoring tool for unauthorized attempt to access the data. As far as HIS working with sensitive personal data, then all data manipulations should be logged to allow handle security incidents. This specific feature of the system will require installing and configuring one of the modern ticketing systems that should be customized to fit the HIS security requirements.

There exist many ticketing systems on the market both commercial or public, that could be customized for this issue [5]. It is reasonable to highlight three suitable types of ticketing systems:

1. Commercial: like “Jira Service Desk” or “Service Now” – the powerful solutions that have multiple integration options. The main disadvantage is that those are not cost effective.
2. Public/Free: the most popular example is Request Tracker (RT), that is free distributed, also has good documentation and customization features.
3. Custom: that could be developed especially for a medical information system. The main advantage of this option is that it could be easily integrated directly with patient medical record.

For all three options listed above incident handling procedure is implemented by the following steps:

Step 1: Receiving incident reports. Incident reports reach via several channels, mostly by e-mail, but also by telephone or on-line messages. Notes are made for all available details in a fixed format while receiving the incident report.

Step 2: Incident evaluation. The authenticity and relevance of the reported incident is verified and the incident is classified (by category, criticality and sensitivity). Triage is on the critical path for understanding what is being reported throughout the organization.

Step 3: Actions. Usually triaged incidents go into a request queue in an incident handling tool that is used by one or more incident handlers.

3.1 Start the incident ticket handle. Create incident ticket number, if it hasn't been created automatically.

3.2 Incident lifecycle. This circle contains the following processes: analysis, obtain contact information, provide technical assistance, coordination.

3.3 Incident handling report.

3.4 Archiving.

One incident could be linked to one or many medical images that offer the traceability for whole process.

4. Conclusion

Information security and personal data protection is a critical functionality for any medical information system. This becomes more and more demanded in the modern world because hospitals and patients require tools for being able to share the data and have access to historic medical records from any place in the world.

Medical images are specific data that need solving many specialized tasks to make sure that the data is protected. In practice, this means that each MIS, HIS and LIS (Laboratory informational system) requires development of additional modules that could on one the hand protect the data and on the other hand make it possible the data sharing based on request/approval mechanism.

All medical data manipulations should be tracked, and in case of any security issues a ticket should be generated and sent to specialists for

analysis. Based on the number of patients and estimated number of incidents it is not possible to handle all the security tasks without optimization and automatization of incidents handling processes. That will require installing a specialized helpdesk that should be integrated with the patient medical records.

The activities described above should make the entire patient treatment process more secure and eliminate possible data leaks.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

References

[1] *DICOM format description* - http://dicom.nema.org/standard.html.

[2] Alexandr Golubev, Peter Bogatencov, Grigore Secrieru, and Nicolai Iliuha. *DICOM Network - Solution for Medical Imagistic Investigations Exchange*. International Workshop on Intelligent Information Systems. Proceedings IIS. 13-14 September, Chisinau, IMI ASM, 2011, pp. 179-182. ISBN 978-9975-4237-0-0.

[3] Peter Bogatencov, Nicolai Iliuha, Grigore Secrieru, and Alexandr Golubev. *DICOM Network for Medical Imagistic Investigations Storage, Access and Processing*. “Networking in Education and Research”, Proceedings of the 11th RoEduNet IEEE International Conference, Sinaia, Romania, January 17-19, 2013, pp. 38-42. ISSN-L 2068-1038.

[4] A. Anagnostaki, S. Pavlopoulos, E. Kyriakou, and D. Koutsouris, *A Novel Codification Scheme Based on the VITAL and DICOM Standards for Telemedicine Applications*, IEEE Transactions on Biomedical Engineering, vol. 49, no. 12, pp. 1399–1411, 2002.

[5] *The Best Help Desk Software for 2020* - https://www.pcmag.com/picks/the-best-help-desk-software.

Alexandr Golubev[1], Petru Bogatencov[2], Ecaterina Matenco[1], Grigore Secrieru[2]

[1]Affiliation/Institution: RENAM Association, Chisinau, Republic of Moldova
E-mail: galex@galex.md
[2]Affiliation/Institution: Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chisinau, Republic of Moldova
E-mail: bogatenc@asm.md

Distributed computing infrastructure for complex applications development

Petru Bogatencov, Grigore Secrieru, Radu Buzatu, Nichita Degteariov

Abstract

Implementation and use of heterogeneous Multi-zone Cloud infrastructure that integrates various types of computing resources are described. It is shown the necessity of the development of computer infrastructures and services that are focused on supporting Open Science initiatives and offering conditions for solving complex problems with high demands of computing resources. Approaches to the deployment of complex cloud infrastructures, their configuration, and administration are presented. The described computing infrastructure has an important role for the research community of Moldova in using the performances of the creating European Open Science Cloud resources and services.

Keywords: cloud computing, e-infrastructure & services, Open Science support tools.

1. Introduction

The importance of the development of computing infrastructure and services for support of open research data accumulation, storage, and processing is permanently increasing. These e-Infrastructures became more and more universal and provide various types of services for operation with research data, including high-performance resources for complex data processing applications development and porting.

Work on the implementation of distributed computing infrastructure in Moldova started in 2007 when the first Agreement on the creation of the MD-GIRD Joint Research Unit Consortium and accompanying Memorandums of Understanding were signed by seven universities and research institutes of Moldova. Since this time, the works started on the

deployment of the first national distributed computing infrastructure that included integration of computing clusters and servers installed in the Technical University, State University, State University of Medicine and Pharmaceutics, Vladimir Andrunachievich Institute of Mathematics and Computer Science, RENAM, and State Hydrometeorological Service.

Since 2013 in the Vladimir Andrunachievich Institute of Mathematics and Computer Science (VA IMCS), the works started on the deployment of virtualized cloud-based computing infrastructure and transferring of the existing in Moldova distributed computing infrastructure to a distributed cloud environment. Participation of VA IMCS and RENAM specialists in the regional project "Experimental Deployment of an Integrated Grid and Cloud-Enabled Environment in BSEC Countries on the Base of gEclipse (BSEC gEclipseGrid)" supported by the Black Sea Economic Cooperation Programme (http://www.blacksea-Cloud.net) in 2014-2015 had a significant impact on support of these works [1]. In continuation (2014-2022), deployment of the national cloud computing infrastructure was supported by the bilateral project "Instrumental support for complex applications porting to the regional HPC infrastructure" funded by the Science and Technology Center in Ukraine and the Academy of Sciences of Moldova [2] and by EU funded projects: Eastern Partnership Connect (EaPConnect), EU4Digital - Connecting Research and Education Communities (EaPConnect2). These projects significantly contributed to the procurement of new computing equipment and cloud computing infrastructure extension. Now, these works are supported by the national project "Investigation and elaboration of the integrated infrastructure of the unified environment "cloud computing" to support open science" funded by the National Agency for Research and Development [3].

2. Approaches for the Distributed Computing Infrastructure Deployment

In the first stage, it is planned to deploy a multi-zone IaaS Cloud infrastructure that combines the resources of VA IMCS, the State University of Moldova (SUM), and RENAM into distributed computing network for processing scientific data, performing intensive scientific calculations, as well as storing and archiving research data and results of computational experiments.

Works on a new Scientific multi-zone IaaS Cloud Infrastructure that is based on OpenStack Ussuri have begun in 2021 and are progressing now. As a result, today in VA IMCS and RENAM, in parallel are available previously deployed and operating resources via cloud.renam.md, which runs on an outdated version of OpenStack Mitaka and a new Cloud platform, based on the latest OpenStack Ussuri version, offering more features, more processing power, and flexibility of operation. During the pilot cloud infrastructure testing and subsequent work on the outdated version of OpenStack Mitaka, many bottlenecks were identified and several ideas were proposed for performance enhancements. There are several improvements to the new platform accessible via openstack.math.md comparing with the old one. In the new platform, now 4 servers are used instead of 1 used previously as a host system with the following parameters: 2 servers with 24 vcpu and 2 servers with 16 vcpu; 48 Gb RAM on each server and 1 TB HDD space for creating virtual machines.

A new important component has been added - block storage, which allows the creation of volumes for organizing persistent storage. In general, in OpenStack, as in other modern Cloud systems, several concepts exist for providing storage resources. When creating a virtual machine, you can choose a predefined flavor, with a predefined number of CPU, RAM, and HDD space; but previously, when you delete a virtual machine, all data stored on the machine instantly disappears. The new storage component, used in the created multi-zone IaaS Cloud Infrastructure, is deployed on a separate storage server and allows you to create block storage devices and mount them on a virtual machine through special drivers over the network. This is a kind of network flash drive that can be mounted to any virtual machine associated with the project, unmounted and remounted to another, etc., and most importantly, this type of volume is persistent storage that can be reused when the virtual machines are deleted. Thus, you can no longer worry about data safety and easily move data from one virtual machine to another, or quickly scale up VM performance by creating a virtual machine with larger resources and simply mount volumes to it with all scientific data available for further processing.

Now, for guest systems, two separate subnets with 32 IP addresses each have been created, as opposed to the one subnet with 16 available IP addresses in the previous version.

A more advanced and flexible model of interaction with the network has been implemented. In the new cloud infrastructure (openstack.math.md), in addition to the usual "provider network" model, which allocates one real IP address from the pool of provider network addresses to each virtual machine, a self-service network is also available. A self-service network allows each project to create its own local network with Internet access via NAT (Network Address Translation). For a Self-service network, the user creates a virtual router for the project with its own address space for the local network. Virtual eXtensible Local Area Network (VXLAN) traffic tagging is used to create such overlay networks that prevent the occurrence of address conflicts between projects in case several projects will use network addresses from the same range. To ensure the functioning of NAT, one IP address from the provider network is allocated to the external interface of the virtual router, which serves as a gateway for virtual machines within the project. Also, when using the self-service model, the floating IP technology becomes available, which allows you to temporarily bind the IP address from the provider network to any of the virtual machines in the project, and at any time detach it and reassign it to any other virtual machine of the project. Moreover, the replacement occurs seamlessly, that is, the address does not change inside the machine, but remains the same - the address is from the internal network of the project, but the changes occur at the level of the virtual router. Incoming to the external address packets are forwarded by the virtual router to the internal interface of the selected virtual machine. This allows you efficiently to use IP addresses and not allocate an external address to each virtual machine. The external IP address remains assigned to the project and can be reused by other machines within the project.

For the deployment of new computing infrastructure, the process of transition to a 10G network has started according to the elaborated plan. The New Juniper switch already has been installed and all storage servers with 10G cards on board have been connected to this switch. We have a plan to switch all remaining servers to 10G interfaces this year. To increase the bandwidth and improve the reliability of the existing 1G

network, Linux bond technology has been applied to the existing network, which allows aggregating two or more network interfaces into one logical device by selecting one of seven possible modes of operation [4]. We use the balance-rr mode, which balances traffic by distributing network packets sequentially from the first interface to the last. This allows getting a twofold or more increase in throughput when combining two or more network interfaces into a bond. We use this technology by combining two 1G interfaces into one 2G interface for connecting compute and storage nodes for faster and more stable operation of a guest OS with persistent volumes (see Fig. 1). As you can see, data is transferred between the two servers with the speed 1.94 Gbps for sending and 1.93 Gbps for receiving.

This temporary solution will remain operational until the existing 1G network infrastructure will be completely switched to 10G.

```
root@pve002:~# iperf3 -c 192.168.0.125
Connecting to host 192.168.0.125, port 5201
[  5] local 192.168.0.127 port 48862 connected to 192.168.0.125 port 5201
[ ID] Interval           Transfer     Bitrate         Retr  Cwnd
[  5]   0.00-1.00   sec   226 MBytes  1.90 Gbits/sec  278    760 KBytes
[  5]   1.00-2.00   sec   234 MBytes  1.97 Gbits/sec  315    690 KBytes
[  5]   2.00-3.00   sec   227 MBytes  1.90 Gbits/sec  640    638 KBytes
[  5]   3.00-4.00   sec   234 MBytes  1.97 Gbits/sec  158    725 KBytes
[  5]   4.00-5.00   sec   224 MBytes  1.88 Gbits/sec  635    166 KBytes
[  5]   5.00-6.00   sec   230 MBytes  1.93 Gbits/sec  251    743 KBytes
[  5]   6.00-7.00   sec   230 MBytes  1.93 Gbits/sec  398    664 KBytes
[  5]   7.00-8.00   sec   236 MBytes  1.98 Gbits/sec   11    909 KBytes
[  5]   8.00-9.00   sec   235 MBytes  1.97 Gbits/sec   19    979 KBytes
[  5]   9.00-10.00  sec   231 MBytes  1.94 Gbits/sec  528    839 KBytes
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bitrate         Retr
[  5]   0.00-10.00  sec  2.25 GBytes  1.94 Gbits/sec  3233             sender
[  5]   0.00-10.00  sec  2.25 GBytes  1.93 Gbits/sec                  receiver

iperf Done.
```

Figure 1. Speed measurement on the bond interface

Another significant improvement is the ability to configure guest machines using cloud-config as shown in Fig. 2. Various parameters and commands can be passed to the guest virtual machine at the boot stage to configure it, which allows you to fully automate OS pre-configuration, package installation, starting services, etc.

In this example, we illustrate creating a virtual machine with pre-configuration parameters passed to the VM that, for example, allows creating new user, defining a password for the user, appending a ssh key for passwordless access, and installing new packages and updates. The last command sends an email with the subject "Your Openstack VM is

ready" and the body text "Now you can ssh to it, cheers!" to the predefined user mail address.

```
#cloud-config
users:
  - name: nikita
    groups: [adm, audio, cdrom, dialout, floppy, video, plugdev, dip, netdev, sudo]
    passwd:
$6$1rj3YMu0hbgxIR$Sm4QjQdN0jYFcD/HcCvP9k1KCZsJ3eiTPEJ6aF7ZoTWAtaG6apsQNz0BT3afhlUDZQDL0Uj.fi0ySBKYn
aGPS1
    ssh_authorized_keys:
      -                                                                                "ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQCj8YYTA+pcq7rZzQd7r8ClOqbkGHABkbAwBrNy2QG+BFwMStp9dg8Ynf9xlJVdqwh8KAX
9kCiJPxSFFH97HqCjfjET1kOBpTI99Bp2R0NfmIH2NALKJgTzQo4mWFLd0AgO82M0vmANQXpF2slRfbPjqGWkJQQRzwm0/YiKAg
kzQU/+Es98i03g9JWyvQXoToqt9NZoLGMLiwh/savlEl163YYf9T+vRmzo2bmHARh5bkGa8RE0Gf6PhK3Z1HdFcOTUdVvtb2Hhx
1XSKFIvj7S7/PqIexjBaU1WWIR59SCuzVnlYQzR+XEh8xEXaiFegBj3Wud9ZFA5t6wAfjOt30O3 nikita@cloud"
  package_upgrade: true
  packages:
    - fail2ban
    - sendmail
    - mailutils
  runcmd:
    - systemctl restart sendmail
    - systemctl restart fail2ban

    -  echo  "Now  you  can  ssh  to  it,  cheers!"  |  mail  -s  "Your  Openstack  VM  is  ready"

nichita.degteariov@math.md
```

Figure 2. Configuration example using cloud-config.

Created robust, performant, and secure computing infrastructure has become a platform for expansion and improvement of services provided for different research teams.

3. Use of the Computing Infrastructure for Complex Applications and Algorithms Realization

Distributed computing infrastructure supports the adaptive execution framework that can be configurated and tested for the solution of different complex applications. The research teams from VA IMCS and SUM developed and continue developing several applications that require resources of multiprocessor clusters and distributed computing infrastructure. One practical example of the application for solving complex decision-making problems elaborated for porting on the created computing infrastructure was described in [3]. Other examples of complex applications and services deployment by using resources of the created computing infrastructure are presented below.

3.1 Optimal Partition of the State into Economic Territorial Units

The main algorithm for determining the optimal territorial partition of the state into coherent economic territorial units (ETUs) was proposed and realized [5]. The algorithm allows solving the problem that involves merging a set of localities having strong infrastructure communication, which comprises telephone lines, gas pipes, road, power, and water systems. So, the resulting ETUs have to meet multiple criteria, which need to be balanced. The criterion, which is hardest to satisfy is the territorial contiguity of an ETU; it essentially means that it is possible to travel by roads between any two localities within the ETU without having to visit locality from other ETUs.

The solution process of the described problem consists of the following three steps:

1) Elaboration of two appropriate integer linear programming models (ILP models), which can be easily adjusted to special restrictions and criteria;
2) Determination, by using the first ILP model, the optimal number of ETUs needed to partition the state, for which the imposed restrictions are satisfied;
3) Determination, by using the second ILP model, the most balanced partition into the optimal number of ETUs established at the previous step.

The contiguity of ETUs in ILP models, used in steps 2) and 3), is expressed by the shortest roads (paths) between the ETUs centers and any other locality assigned to these centers. This formulation substantially reduces the number of variables and constraints used in the ILP models and, as a result, a balance between the quality of the solution and the computational effort is achieved. We consider that the maximum allowable travel distance D from the center of an ETU to any commune assigned to it is the main restriction of the elaborated IPL models. Depending on the choice of D, the algorithm will determine the optimal partition.

The developed algorithms, essentially consisting of steps 2) and 3), will be implemented and tested on the created distributed computing infrastructure that includes clusters with multicore processors by means of Python programming language and CPLEX Optimization Studio.

3.2 DICOM Network - Distributed System for Medical Images Pre-processing and Archiving

The "DICOM Network" project was launched in Moldova in 2012, whose goal is to provide access to the collected imagistic data for medical staff with the appropriate access rights and for patients - to the personal radiography investigations. Today the system implemented in many hospitals in Moldova, collects and processes more than 5TB of data per month gathered from different types of medical equipment [6].

"DICOM Network" realization based on the national scientific cloud platform opens many possibilities for using this application for various types of activities. DICOM investigations could be added to some other datasets, that are collected and available from cloud infrastructure. But additional functionalities require the realization of supplementary solutions for Imagistic Data anonymizing [7]. Cloud technologies and services allow optimization and making collected medical imagistic data compatible with Open Science principles; they become widely accessible by means of mobile devices.

The diagram below (see Fig. 3) shows different data processing options that make it possible to store collected data in cloud storage.

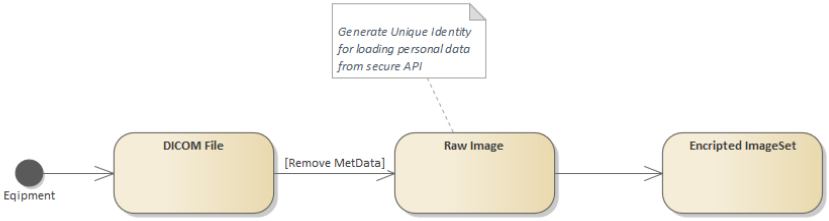


Figure 3. Data processing options for storing data in the cloud storage.

Initially, all data is collected in DICOM (.dcm) format that contains raw images and XML data with personal information. The proposed data format optimizations could be divided into two steps:

The first is the removal of XML with personal data from the image file. The extracted data will be sent as metadata and stored under a unique identifier for this image set study. Anonymized this way data can be stored in the local network of the data owner.

The second is encrypting the full image set, that will completely eliminate the possibility of restoring the personal data based on the image. This option will make it possible to store encrypted images using external cloud storage facilities.

This will require additional image processing for data storage, data accessing, and data visualization, but it makes all personal data protected. At the same time, this approach makes it possible to use various cloud-based APIs and services for data processing and exchange.

3.3 Integrated system for distant learning and video-conferences support

Since 2019, the created Cloud infrastructure started to use for a new service – support of online lectures organization for the State University and the Technical University of Moldova. These realizations allowed to intensify the use of distributed computing resources. One of the representative examples is launched in 2020 and actively used - a multi-node distributed video-conferencing system, that provided facilities for the organization of online classes since the beginning of the lockdowns, caused by the COVID-19 pandemic back in 2020. The video-conferencing system is powered by the open-source project BigBlueButton [8]. The system is integrated with Moodle, creating a self-sufficient distant e-learning platform and it is actively used for distant learning by the main universities of Moldova: the Moldova State University, the Academy of Economic Studies of Moldova, as well as by some smaller institutions in Chisinau and the regions (e.g., in Comrat and Taraclia cities). It hosts roughly 1 – 1,2k concurrent users daily with peaks up to nearly 2k users in about 60 separate virtual rooms distributed among the servers' cluster. As an example, the statistics for the first half of September 2021 are presented in Fig. 4.

The effective use of the VC system has been achieved by uniting distributed BBB nodes in a cluster using the Scalelite project. Scalelite is an open-source load balancer that manages a pool of BBB servers. It makes the pool of servers appear as a single (very scalable) BigBlueButton server. A front-end, such as Moodle or Greenlight platforms, sends standard BBB API requests to the Scalelite server which, in turn, distributes those requests to the least loaded BigBlueButton server in the pool. We also use Greenlight as a meeting managing plugin and a

pool of three Traversal Using Relay NAT (TURN) servers for relaying the traffic between peers behind the NAT.

Figure 4. Number of VC system users

5. Conclusion

With the support of several national and international projects, the distributed computing infrastructure in Moldova was created, jointly operating by interested institutions, and permanently upgraded by the installation of new computing equipment and new versions of middleware. The created computing infrastructure ensures reliable operation and wide access to its resources, deployment of various tools and platforms for support of Open Science that corresponds to the current needs of researchers in Moldova. Deployment of modern architectural solutions, tools, and platforms, and installation of new high-performance servers allowed intensifying the use of the distributed computing resources and contributed to the qualitative development of IT services for R&E in Moldova.

Acknowledgments. Grants from the National Agency for Science and Development (grant No. 20.80009.5007.22 and grant No. 21.70105.9ŞD) and EU funded "EU4Digital: Connecting Research and Education Communities (EaPConnect2)" project (grant contract ENI/2019/407-452) have supported part of the research for this paper.

References

[1] Nichita Degteariov, Petru Bogatencov, Nicolai Iliuha, and Grigore Secrieru. *Upgrading Cloud Infrastructure for Research Activities Support*. Workshop on Intelligent Information Systems (WIIS2020) Proceedings, Chisinau, IMI, 2020, pp. 69-74.

[2] P. Bogatencov, N. Iliuha, E. Calmis, B. Hancu, V. Patiuc, and G. Secrieru. *Computational Infrastructure for Porting and Execution Complex Applications in Moldova.* Proceeding of the 5th International Conference "Telecommunications, Electronics and Informatics", May 20 – 23, 2015, Chisinau, UTM, 2015, pp. 21-26.

[3] Petru Bogatencov, Grigore Secrieru, Boris Hîncu, and Nichita Degteariov. *Development of computing infrastructure for support of Open Science in Moldova.* Workshop on Intelligent Information Systems (WIIS2021) Proceedings, Chisinau, IMI, 2020, pp. 34-45.

[4] *Link aggregation*, https://en.wikipedia.org/wiki/Link_aggregation.

[5] Radu Buzatu and Sergiu Cataranciuc. *On Nontrivial Covers and Partitions of Graphs by Convex Sets.* Comput. Sci. J. Moldova 26(1), IMI, 2018, pp. 3-14.

[6] Peter Bogatencov, Nicolai Iliuha, Grigore Secrieru, and Alexandr Golubev. *DICOM Network for Medical Imagistic Investigations Storage, Access and Processing.* "Networking in Education and Research", Proceedings of the 11th RoEduNet IEEE International Conference, Sinaia, Romania, January 17-19, 2013, pp. 38-42.

[7] A. Anagnostaki, S. Pavlopoulos, E. Kyriakou, and D. Koutsouris. *A Novel Codification Scheme Based on the VITAL and DICOM Standards for Telemedicine Applications*. IEEE Transactions on Biomedical Engineering, vol. 49, no. 12, pp. 1399–1411, 2002.

[8] Grigore Secrieru, Peter Bogatencov, and Nichita Degteariov. *Development of Effective Access to the Distributed Scientific and Educational e-Infrastructure*. Proceedings of the 9th International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID'2021). Dubna, Russia, 5-9 July 2021; Vol-3041, urn:nbn:de:0074-3041-3, pp. 503-507; DOI:10.54546/MLIT.2021.21.73.001.

Petru Bogatencov[1,2], Grigore Secrieru[1], Radu Buzatu[3], Nichita Degteariov[1,2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science. Chisinau, Moldova.
E-mail: bogatencov@renam.md, secrieru@renam.md, nichita.degteariov@math.md

[2]RENAM Association. Chisinau, Moldova.
E-mail: bogatencov@renam.md, nichita.degteariov@math.md

[3]State University of Moldova. Chisinau, Moldova.
E-mail: radubuzatu@gmail.com

Development of computing infrastructure for support of Open Science in Moldova

Petru Bogatencov, Grigore Secrieru,
Boris Hîncu, Nichita Degteariov

Abstract

In the paper, there are considered the e-Infrastructures which enable instruments and provide facilities, resources, and services that are used by the research communities to conduct research and foster innovation in their fields. The European Open Science Cloud (EOSC) will be the open and trusted virtual environment which will allow European researchers to store, share and reuse research data across borders and disciplines. Elaboration of the distributed computing infrastructure to support the Open Science initiative in Moldova has an important role for the research community in using the performances of the EOSC ecosystem.

Keywords: computing infrastructure, research infrastructures, e-Infrastructures, EOSC ecosystem, platforms, and tools for Open Science.

1. Introduction

The exponential growth of information and the availability of digital technologies have generated a new approach to scientific research - Open Science, which is based on accessibility, collaboration, e-infrastructures, and new ways of disseminating knowledge. The following key circumstances are associated with the development of the infrastructure for Open Science support:

- Most of the existing infrastructure of Open Science, in addition to open access to publications, can be considered to be at the beginning of the path;

- We need a stable, reliable, and evolving infrastructure that will support the creation of more research data, with the possibility of their reuse, as well as digital research tools that are efficient and easy to use,
- Some elements of the global open science infrastructure require collective efforts for their creation and management.

These circumstances aim to achieve a consensus on ensuring equitable access to the advantages of Open Science and participation in its development.

The main support for the development of Open Science is provided by the EOSC ecosystem [1], which will be developed in the period 2021-2027 within the Horizon Europe research-innovation program. EOSC constitutes a major ambition in the European Open Science policy, being a federated ecosystem of research infrastructures, e-infrastructures, and services that allow the scientific community to share and process publicly funded research results and data across borders and scientific domains.

The National Open Science Cloud Initiatives (NOSCI) are important pillars of this effort [2]. For ensuring the sustainability of this flagship initiative for Europe, it has to be built upon solid governance and organizational framework on the national level.

In order to stimulate the participation of countries from different regions of Europe in the promotion, expansion, and use of the resources of the EOSC ecosystem in progress are European projects, including the project “NI4OS Europe – National Initiatives for Open Science in Europe”. The NI4OS-Europe project proposes a modular workflow for the integration of national resources and services into EOSC and the setup of NOSCI. RENAM Association, as a participant in the NI4OS-Europe, initiated, according to the project requirements, the process of establishing the national initiative on the Cloud for Open Science – MD-NOSCI, developed "NOSCI Establishment Roadmap for the Republic of Moldova", presented to the Ministry of Education and Research. In this way, RENAM provides the Moldovan research community with information and access to up-to-date resources, representing the main point of contact with international initiatives in the field of open science. NOSCIs in the Member States and partner countries have an important

role to play in using performance and facilitating EOSC governance. Fig. 1 expresses an overview of the EOSC ecosystem, presented in [1].

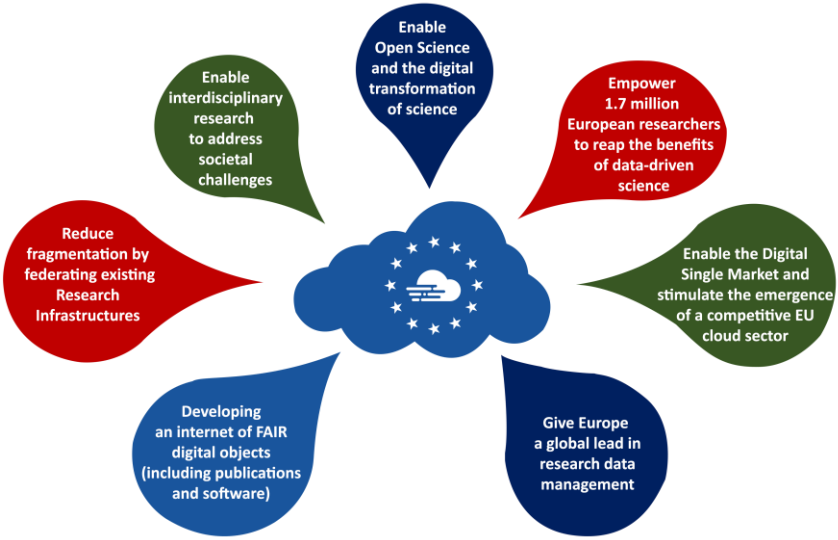


Figure 1. EOSC vision.

NOSCI can be seen as a coalition of national organizations that have a prominent role and interest in the European Open Science Cloud. The main goal of NOSCI is to promote synergies at the national level and to optimize their participation in European and global challenges in the field of open science, the development of the EOSC ecosystem that includes as an example the elaboration of the software infrastructure for the cloud-like high-performance computing oriented towards classes of problems in the field of mathematical modeling of decision-making processes.

The details of the NOSCI establishment objectives and the necessary activities can be provided in the Memorandum of Understanding that will be the basis of this initiative.

The State University of Moldova (SUM) and the RENAM share the goal of developing and providing an innovative federated Cloud environment with the efficient use of hard and soft resources available to promote the concept of Open Science, development of digital technologies

in research and education (R&E) and solving problems that require high-performance computing (HPC) and generating large volumes of data.

For the interconnection of universities and research institutes, the RENAM-GEANT networking platform is used at speeds up to 10 Gbps and more based on the national fiber optic backbone of NREN RENAM. These actions are aimed at creating the necessary conditions for the academic and research community to use the performance of the EOSC ecosystem.

2. Development of distributed computing infrastructure

Nowadays, almost any organization has its own IT structure, which unites servers and computers of employees into a common network, or can rent cloud services and refuse from owning physical servers. In describing cloud solutions, there are often three cloud service models that have the following abbreviations: IaaS, PaaS, and SaaS.

Infrastructure as a Service (IaaS) is a solution for hosting infrastructure in the cloud by renting cloud servers. This means creating an infrastructure in the cloud with the required configuration of computing resources in the form of CPU, RAM, HDD, which the customer uses at his own discretion. It is an alternative to investing in expensive hardware and networking equipment. IaaS allows you to organize a pool of virtual machines, prepare remote workstations, data storage, etc., providing quick access to reliable, secure, flexible, and scalable infrastructure.

The IaaS configuration model for computing resources, hardware, and network equipment includes:

1. Virtual servers (VPS / VDS) on which you can install various software products. The provider can offer servers with operating systems so that you can quickly deploy the necessary applications to them.
2. Network links that allow virtual servers to communicate with each other, external servers owned by the customer, and the Internet. These include:
    - server availability for each other and the external network, routing of server network connections;
    - load balancing, which prevents server overload by distributing incoming traffic between the pool of servers;

- VPN – technology for encrypting data transmitted by a company between the cloud and its physical data center;

3. User access control. For example, you can restrict access to individual virtual machines or allow viewing of data, but prohibit making changes to them.
4. Cloud storage for storing files, data, or backups. They differ from ordinary cloud drives, which individual users deal with, with almost unlimited storage capacity and high speed of data access.
5. Backup services and resilience disasters that insure infrastructure against falls and data loss in the event of failure of its individual nodes.

The Platform as a Service (PaaS) and the main differences from the IaaS model. In the case of PaaS, the customer is provided with certain tools such as a database management system, a big data processing environment, which need to be customized to the needs of the organization, but do not need to be built from scratch. At the same time, there is no access to the operating system and the settings of virtual servers that underlie PaaS, there is only access to the interfaces of the platform itself. In the case of IaaS, we only get disk space and must choose an open research data management system (ORDMS), install and configure it, ensure data protection and backup. In PaaS, the ORDMS is already installed, you just need to configure it for yourself and manage your data.

The Software as a Service (SaaS) model is a fully configured, out-of-the-box software environment that performs specific functions. The software itself is in the cloud and is accessed via a network, and the software environment runs on the capacities of virtual servers. Most of the services on the Internet can serve as examples of SaaS: email, task schedulers, web builders for creating sites, open research data repositories, and other cloud applications for solving specific problems.

SUM, Vladimir Andrunachievici Institute of Mathematics and Computer Science (IMCS), and RENAM jointly realize the goal of developing and providing resources of the innovative federated Cloud environment that ensures the efficient use of available hardware and software resources to promote the concept of Open Science, development

of digital technologies for research and education (R&E) and solving problems that require high-performance computing resources and processing of large amounts of data.

The key points for the creation and development of a distributed infrastructure to support open science were facilitated by the participation of SUM, RENAM, and IMCS in international and national projects that allowed developing and modernizing the computing hardware and network equipment: SUM, RENAM, and IMCS computing clusters were upgraded, new high-performance servers for compute nodes and storage elements were purchased; 10 Gbps optical network deployed between the main computing locations; IMCS and central RENAM nodes were modernized – modern uninterruptible power supply systems for server equipment and industrial air cooling systems were installed.

Current activities focused on deployment and development of HPC cloud infrastructure by implementing SaaS, PaaS, and IaaS service models for the joint use of parallel computing clusters of SUM, IMCS, and RENAM integrated into a single distributed Cloud environment. Thus, a cloud service will be created that is focused on performing various types of tasks, hosting ORDMS, and is suitable for deploying open science tools and services.

Until now Cloud IaaS based on OpenStack [3] is running on the IMCS-RENAM computing infrastructure and has been available for scientists for several years. The system is successfully used by researchers in several resource-intensive projects that use Machine-Learning technologies to train neural networks in text recognition, language processing, etc. The existing system has some disadvantages due to the limited number of available resources and lack of a high throughput network interconnection between nodes and storage elements [4]. The creation of the new joint SUM-RENAM-IMCS Cloud infrastructure is aimed at solving these known problems.

3. Expected results and future plans

At the first stage, it is planned to create a multi-zone IaaS Cloud infrastructure that combines the resources of IMCS, SUM, and RENAM into a distributed computing network for processing scientific data,

performing scientific calculations, as well as storing and archiving results (see Fig. 2).

PaaS and SaaS solutions will be deployed on the created computing infrastructure: tools for processing and analyzing scientific data, such as Juniper Notebook, Python, Elastic Search, and others; open repositories for scientific data collection will be installed. The tools and platforms proposed for deployment will be integrated (onboarded) into the EOSC Portal.

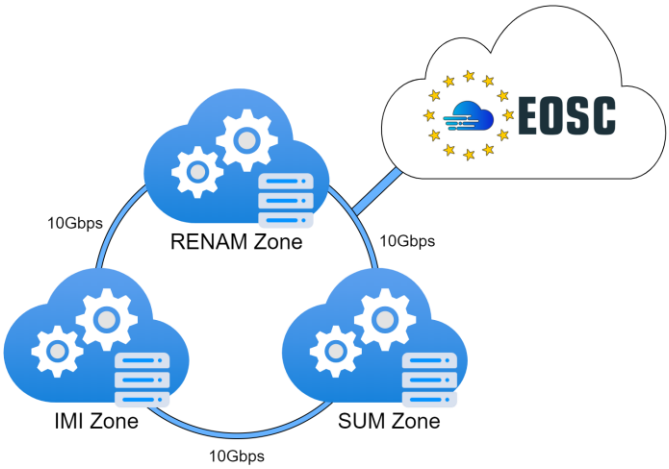


Figure 2. Multi-zone Cloud: IMI – RENAM – SUM.

Integration of new servers will increase the computing power, and the combination of computing power and data storage systems of the three organizations will allow more flexible and efficient use of computing resources for solving large-scale problems. Parameters of the newly created distributed computing infrastructure:

- IMCS Compute servers: Dell R540 servers with total of 32 CPU cores, 128GB RAM, Dell R740 server with 12 TB of RAID of storage;
- RENAM Compute servers: Dell R730 servers with total of 40 CPU cores, 256 GB RAM, Dell R 740 server with 12 TB of RAID storage;
- SUM Compute servers: HP ProLiant DL140 G3 with total of 28 CPU cores, 82 GB RAM, 2TB of RAID Storage.

An analysis of the hardware and software resources of the parallel clusters of SUM, RENAM, and IMCS will be carried out with the aim of coordinated and safe use of joint resources. Optimal configurations of hardware and software resources of available clusters and servers will be elaborated and realized for effective use of the available resources in the Cloud infrastructure.

Distributed computing infrastructure will use the adaptive execution framework that can be adapted and tested for the solution of different complex applications. The research team from SUM developed several applications that require resources of multiprocessor clusters and distributed computing infrastructure. One practical example of the application for solving complex decision-making problems – elaborating for porting on the created computing infrastructure – is described below.

3.1 The parallel algorithm for determining Nash equilibrium profiles in modeling decision-making processes.

As mentioned above the developed Cloud HPC computing system can be used for operating data storage and analysis platforms, executing various scientific applications, including modeling decision-making processes in situations of risk, conflict, and informational impact, using the mathematical apparatus of game theory. In this paragraph, we will describe and analyze how to test the Cloud HPC system by running parallel programs. As an example, the parallel algorithm for determining Nash equilibrium situations in modeling decision-making processes is proposed.

Contemporary decision-making problems are very complex and require the processing of a very large volume of data. Thus, for the mathematical modeling of these processes, it is necessary to take into account the big data problems. Big data is a huge amount of data that is beyond the processing capacity to manage and analyze the data in a specific time interval. The data is too big to be stored and processed by a single machine. In many large-scale solutions, data is divided into partitions that can be managed and accessed separately. In order to solve such problems in real-time, parallel algorithms are built and then implemented on various types of parallel computing systems. For parallel data processing, we must use the ways of dividing, partitioning (sharing), and distributing data. Different paradigms and programming models can

be used for the soft implementation of parallel algorithms on distributed memory computing systems (parallel clusters, cloud computing systems). From the multitude of parallel programming models here we will present how to implement software on HPC clusters of the model based on MPI functions.

For modeling the decision-making problems we will consider the bimatrix game in the following strategic form $\Gamma = \langle I, J, A, B \rangle$, where *I={1,2,..,n}* is the line index set (the set of strategies of the player 1), *J={1,2,..,m}* is the column index set (the set of strategies of the player 2) and $A = \|a_{ij}\|_{\substack{i \in I \\ j \in J}}$, $B = \|b_{ij}\|_{\substack{i \in I \\ j \in J}}$ are the payoff matrices of player 1 and player 2, respectively. We denote by *NE[Γ]* the set of all equilibrium profiles in the game *Γ*. Thus, the Nash equilibrium profile is the pair of indices (i^*, j^*), for which the following system of inequalities is verified $(i^*, j^*) \Leftrightarrow \begin{cases} a_{i^* j^*} \geq a_{ij^*} \ \forall i \in I, \\ b_{i^* j^*} \geq b_{i^* j} \ \forall j \in J. \end{cases}$ Based on this definition, it is easy to develop the parallel algorithm for determining the Nash equilibrium profiles in bimatrix games.

The structure of the parallel algorithm is determined by the parallelization mode at the data level. That is, the following ways of dividing and distributing matrices A and B can be used:

- Matrices are divided into rectangular submatrices of any size. In this case, the way of constructing equilibrium profiles for the game with the initial matrices is very complicated;
- Matrices are divided into line-type submatrices or column-type submatrices. In this case, building the equilibrium profiles for the initial game is quite simple.

We will mathematically describe the parallel algorithm for determining Nash equilibrium profiless in pure strategies for the bimatrix game defined above. We will assume that matrix A is divided into column-type submatrices and matrix *B* is divided into row-type submatrices. So, we're going to get a series of submatrices $SubA^t = \|a_{ij}\|_{\substack{i \in I \\ }}^{j \in J_k}$ and $SubB^t = \|b_{ij}\|_{i \in I_k}^{j \in J}$, where $I_k = \{i_k, i_{k+1}, ..., i_{k+p}\}$ and

$J_k = \{j_k, j_{k+1}, ..., j_{k+p}\}$. $SubA^t$ is a submatrix that consists of p columns of matrix A starting with column number k and is "distributed" to the process with the rank t. Similarly, SubBt is a submatrix that consists of p lines of the matrix B starting with the line k and is also distributed to the process with the rank t. Using the sequential algorithm described above, the process with the rank t will determine a graph of the point-to-set application $i^*(j_k) = Arg\max_{i \in I} a_{ij_k}$. for any $j_k \in J_k$. Similarly, the process with the rank t will determine a graph of the point-to-set application $j^*(i_k) = Arg\max_{j \in J} b_{i_k j}$. for any $i_k \in I_k$. Finally, the process with rank t will determine $LineGr^t = \bigcup_k gr_k i^*, ColGr^t = \bigcup_k gr_k j^*$. So the Nash equilibrium profiles will be the intersections of the set $\left(\bigcup_t LineGr^t\right)$ and $\left(\bigcup_t ColGr^t\right)$.

Using the MPI parallel programming model [5] on the parallel computing system with distributed memory, a parallel program was developed and tested on control examples to determine Nash equilibrium profiles in bimatrix games. The results of the calculations are presented in Table 1:

Table 1. Computation time for determining solutions in two-matrix games.

Dimensions of the matrices	Runing time (seconds)			
	16 processors	28 processors	32 processors	48 processors
n=30000 m=30000	11.665889	11.187384	14.855256	14.361466
n=35000 m=35000	18.748501	12.359752	13.841478	10.827413
n=40000 m=40000	17.021988	12.191731	13.195590	11.202775
n=45000 m=45000	33.997708	14.808254	13.028965	12.560618
n=48000 m=48000	69.756550	20.810968	18.946320	15.717546

n=49000 m=49000	110.552440	32.880968	16.161673	15.140105
n=49500 m=49500	186.125029	40.321999	22.643520	17.640198
n=49900 m=49900	eror	55.035883	17.819622	19.017421
n=50000 m=50000	eror	46.226555	16.388920	11.839626
n=55000 m=55000	eror	eror	57.292023	38.964008
n=60000 m=60000	eror	eror	eror	156.085372

We have to mention that the developed application we plan to use for testing efficiency of the developed distributed HPC Cloud system, i.e. to use it as a benchmarking program.

4. Conclusion

At the European level, the main support for the development of open science is provided by the EOSC ecosystem, which will be developed in the period 2021-2027 within the Horizon Europe research-innovation program. The possible way to activate participation of the national research community in the European Open Science initiatives is to deploy integrated to EOSC national e-Infrastructures and services. The proposed approach of creation of joint infrastructure and resources for support of Open Science at the national level is one important pave in this direction.

As a principal organizational mechanism that will support the EOSC integration activities is the establishment of the National Initiative MD-NOSCI in Moldova to comply with the European and global challenges in the field of open science. National initiatives in the EU Member States and partner countries play an important role for the academic and research communities in the development and use of open research data and services, as well as in facilitating the governance of the EOSC.

Acknowledgments. This work was supported by the National Agency for Science and Development (grant no. 20.80009.5007.22 and grant No.

20.80009.5007.13) and by the European Commission, project H2020 NI4OS-Europe (grant No. 857645).

References

[1] *EOSC* : https://www.eosc.eu/

[2] *NI4OS* : https://ni4os.eu/

[3] *Cloud IMI-RENAM*: https://cloud.renam.md/

[4] Nichita Degteariov, Petru Bogatencov, Nicolai Iliuha, and Grigore Secrieru. *Upgrading Cloud Infrastructure for Research Activities Support*. In: Workshop on Intelligent Information Systems (WIIS2020) Proceedings, Chisinau, IMI, 2020, pp. 69-74.

[5] B. Hîncu and E. Calmîş. *Modele de programare paralelă pe clustere. Partea I. Programare* MPI. Note de curs. Chişinău: CEP USM, 2016, 129 p.

Petru Bogatencov[1,3], Grigore Secrieru[1,3], Boris Hâncu[2], Nichita Degteariov[3]

[1]PhD/Vladimir Andrunachievici Institute of Mathematics and Computer Science. Chisinau, Moldova.
E-mail: bogatencov@renam.md, secrieru@renam.md

[2]PhD/State University of Moldova. Chisinau, Moldova.
E-mail: boris.hancu@gmail.com

[3]RENAM Association. Chisinau, Moldova.
E-mail: nichita.degteariov@renam.md

Contents

Part 2. Platform for the digitization of heterogeneous documents

Part 3. Intelligent information system structures, databases, and knowledge bases for medical triage and diagnostic applications

Part 4. Automatic content generation systems for computer-assisted training

Part 5. Systemic concept of the heterogeneous multi-cloud platform and methods of realizing the execution environment of imaging information processing applications

Firma poligrafică „VALINEX" SRL, Chişinău, str. Florilor, 30/1A, 26B, tel./fax 43-03-91, e-mail: info@valinex.md, http:\\www.valinex.md

Coli editoriale 12,39. Coli de tipar conv. 14,82. Format 60x84 1/16. Garnitură „Times". Hirtie ofset. Tirajul 150.