Dusan Katuscak – Martin Konvit *

OPTIMISATION OF THE PROCESSES OF WRITTEN HERITAGE PRESERVATION AND DIGITISATION

The Memory of Slovakia Centre of Excellence (MSCOE) is a project of the University of Zilina, under the scientific and organisational auspices of the Faculty of Humanities implemented from 2010 until 2013. The paper contains information about the project, some outcomes and specialisation of scientific and innovation activities. It describes priorities for research and innovation in the MSCOE, especially optimisation of mass digitisation technology, integrating processes of mass digitisation and conservation of textual materials and optimisation of linguistic solutions in mass digitisation and the best possible preparation of digital content for use.

Keywords: *Mass Digitisation, Mass Conservation, Mass Deacidification, Mass Sterilisation, Memory of Slovakia Project, Text Analysis, Knowledge Mining.*

1. The Memory of Slovakia Centre of Excellence

The predominant orientation of Centres of Excellence (COE) in Slovakia since 2008 has been research into new technologies and procedures. The COEs which have been founded are focused mainly on the following areas: technology, medicine, biology, ecology, theoretical computer science, veterinary medicine, social sciences (linguistics).

Among all COE projects approved between 2008 and 2010 the "Memory of Slovakia" Centre of Excellence (MSCOE) is the only one which is primarily focused on the documentation and preservation of cultural heritage. Besides pursuing the general mission of Centres of Excellence, such uniqueness dictates the need for a specific approach to formation of the centre, and formulation of its basic documents as well as profiling the priorities for research, development and innovation in the context of European efforts to promote Europe's cultural and creative industries and the European Commission's Digital Agenda [1], [2], [3].

The Memory of Slovakia – the National Centre of Excellence for Research, Preservation and Accessibility of Cultural and Scientific Heritage (MSCOE) is a project of the University of Zilina. The Slovak National Library in Martin is a partner of the project. The Project is being implemented from September 9th, 2010, until August 31st, 2013 (36 months). The project budget is 4,033,864.35 Eur. On the part of the University of Zilina, the Memory of Slovakia project is supervised by the Faculty of Humanities, the Department of Mediamatics and Cultural Heritage, and the Slovak National Library in Martin is the project partner.

We expect that in 2013 the partners involved will consider various alternatives for integrating the Memory of Slovakia COE and the infrastructure procured into the system of research and development components of the University of Zilina, as well as the potential of research, innovation and organisation of science represented by the MSCOE.

The strategic goal of the project is create the MSCOE for research, preservation and access of cultural and scientific heritage with internationally recognised scientific results. Specific goals of the project are: a) to create the centre of excellence and its formal structure, b) to build ICT infrastructure and complete instrumental equipment of the centre, c) to achieve important scientific results in research, preservation a presentation of cultural and scientific heritage.

2. Strategic Priorities of the MSCOE

1. Identification of the best available technologies in the area of digitisation, optimisation of the methods and tools for mass digitisation, quality control of digitisation, optimisation of transfer of huge amount of data, long time data preservation.
2. Identification of the best available technologies in the area of mass conservation of the books and archive material and integration of processes of mass digitisation and mass conservation of the written heritage.
3. Optimisation of the access to the digital content, creation of the experimental base for availability of scientific and educational content.

* **Dusan Katuscak, Martin Konvit**
Faculty of Humanities, University of Zilina, Slovakia, E-mail: dusan.katuscak@fhv.uniza.sk

Overview of implementation of the project activities Table 1

Item number	Activity name	Progress expressed in %
Main activities		
1.1	Establishment of the Centre of Excellence for research, preservation and presentation of cultural heritage	100%
1.2	Creation of strategy for long-term sustainability of the Centre, coordination of activities, dissemination of the research results	80%
2.1	Procurement, installation and commissioning of the instruments and equipment for the Centre	90%
2.2	Procurement, installation and commissioning of ICT	90%
3.1	Research and development of techniques for access to digital content and techniques of scientific communication	50%
3.2	Research in the area of restoration, conservation and preservation of cultural heritage objects	65%
3.3	Basic research of historical books and collections	70%
Support activities		
Project management		65%
Publicity and awareness-raising		65%

4. Building of the experimental mobile data centre (multi CPU HPC cluster, tiered storage (FC, SATA, Tape), high speed SAN and LAN environment.
5. Research and development of techniques for access of digital content and scientific communication.
6. Research in the area of restoration, conservation and preservation of cultural heritage objects.
7. Basic research of historical books and collections.

3. Implementation of the project activities and available infrastructure (10/2012)

The overview of the project activities implementation is in following Table 1.

Important preliminary results of infrastructure building are illustrated in the next tables: overview of costs in the years 2010 – 2012 [Table 2], mobile data centre components [Table 3], software available for the project purpose [Table 4], installed scanners [Table 5], and list of expected specialised equipment [Table 6].

Overview of costs Table 2

Item	Price
Total project price	3 486 950 €
Mobile data centre	2 735 000 €
Scanners	252 975 €
Specialised equipment	201 100 €
Common ICT equipment	137 750 €
Other	160 125 €

Mobile data centre Table 3

Blade server	• IBM Blade server • 16 blades (32 processors) • divided into 2 physical blocks, each by 8 blades • 2x LAN port (metallic) • 2x SAN port (optic)
Tape library	• robotic, high speed • 4 x LTO5 tape drives with optical connection • 600 LTO5 tapes included • total capacity cca 900 TB
Disc array	• SAN Pillar axiom storage system • 9,6TB - high speed FC discs • 32TB – mass storage SATA discs • HW RAID 0,1,5,6
Hierarchical storage management (HSM)	• incremental backup management • synthetic backup management • VMware and vStorage integration • Backup management thru SAN and LAN
Connectivity	• metallic LAN • optic LAN

Software Table 4

VmWare vSphere	• virtualisation software for Blade server
RedHat Enterprise Linux	• primary operating system
Windows 2008 Server	• secondary operating system
ScanFlow Advanced (ScanGate)	• digitalization workflow software
MediaInfo MIRV	• digital content publishing software
ABBY OCR server	• server-based OCR solution

Scanners Table 5

Treventus	• bound books automated scanning • V-type scanner with book cradle • up to 2 000 pages per hour
BookEye	• bound book binding • flatbed manual scanner • up to 1 500 pages per hour
XinoScan	• simple pages • up to 300 pages per minute

Specialised equipment Table 6

Lux meter	• light intensity measurement
Digital stereomicroscope	• physicochemical properties research of selected materials carriers
Spectrometer	• paper whiteness changes measurement • paper color changes measurement
XRF analyser	• multi elemental analysis of elements from Mg up to U • qualitative analysis of paper document elements • quantitative analysis of paper document elements
SurveNir system	• non-destructive measurement of selected paper document properties

4. Future strategy and perspectives of MSCOE

The priority of the MSCOE in 2013 and the following years is to establish a worksite with expertise in the areas of optimisation of digitisation processes, conservation, preservation, text analysis and knowledge mining.

The MSCOE project is constructed in such a way that it should support the major areas of research in the field of library and information science, mediamatics, mediology and cultural heritage. In accordance with that and with the main research directions the work plan will be organised along these areas of R&D in the following research topics:

Topic 1 is aimed at optimisation of mass digitisation technology

Topic 2 is aimed at integrating processes of mass digitisation and conservation of textual materials.

Topic 3 is aimed at optimisation of linguistic solutions in mass digitisation and the best possible preparation of digital content for use.

5. Area of optimisation of mass digitisation processes (Topic 1)

During the last years, researchers in MSCOE participated significantly in the application of knowledge in practice. By linking the academic sector and one of the leading national memory institutions – the Slovak National Library (SNL) in Martin it was achieved that the SNL started building capacities and infrastructure for mass digitisation of written and printed cultural heritage. Within the relationship among the MSCOE and the National Library and other institutions and systems in the fields of science, research, culture and education, the MSCOE represents a research and experimental base. It is a platform for searching and testing the best available solutions. The combination of research and innovation activities under MSCOE was of key significance especially for the National Library.

The SNL has been awarded the national project (DL&DA) financed from EU structural funds and implemented in 2012 – 2015, with a budget of 49.6 M € and going to produce over 2.8 mil. digitised library & archival objects which amounts to over 270 mil. pages to be selected, treated in mass sterilisation & deacidification, scanned, processed digitally, including image treatment, OCR etc. Within the DL&DA project, the SNL has to create 78 new positions, including researchers, technicians, chemical technologists, mass digitisation and conservation specialists etc. The DL&DA project in figures: Daily production of 43 TB (terabyte), transfer 6 GB (gigabyte) of data per second, digitisation of 2 800 000 objects (270 000 000 pages, represent 17 PB of data, 2 working copies require about 34 PB (petabytes) of data, daily production of 43 TB (terabyte) and need to transfer 6 GB (gigabyte) of data per second [4].

The objectives of Topic 1: support of the research in the field of mass digitisation and digital content reuse through exchange of know-how and experience with partners; recruitment of experienced researchers; specialist dissemination and outreach to innovation capacity building activities. Affordability, widespread availability of tools and services for releasing the economic potential of written library and archival cultural heritage in a digital form and for adding value to the cultural content in an educational, scientific and leisure context. The objectives also include a wider range of users of cultural resources in diverse real and virtual contexts, as well as considerably altered ways of experiencing culture in more personalised and adaptive interactive settings.

All processes of mass digitisation have to be constantly optimised in order to a) increase performance, b) reduce costs, c) identify critical points, d) ensure sustainability, e) improve availability of digital content to people.

Research and similar activities in the DL&DA project are not supported, and, therefore, the optimisation activities must be addressed with regard to the best practices in the MSCOE Project which represents fulfilment of the strategy for preservation and accessibility of documents held in libraries and archives, but not exploited sufficiently for the need of Slovak and European citizens in the interest of overall economic and cultural growth.

The methodology and management of digitisation technology is based on two essential systems: the logistic system which monitors and controls the flow of analogue documents in the digitisation process, and the work flow engine, which is a management system designed on the basis of the SOA/BPM principles, which controls the digitisation itself, interacts with the logistic system,

staff, as well as each automated technological step. Technologically, the digitisation process is supported by cutting-edge HW technology (IB, disk arrays, strong CPU processing), which is technologically prepared for handling large amounts of data. Our calculations show that the financial costs of digitising one page will be 0.20 EUR, which includes complete cleaning, chemical treatment, logistics, technologies, overhead, staff, infrastructures, sterilisation and digital content preservation.

The project's ambition is to demonstrate that these costs are essentially lower than those in similar projects in the EU and the USA. Despite the positive figures indicated, we still see possibilities for optimising the entire process. It has been proven through simple simulations of technological processes that for the given large amount of content to be digitised any minor enhancement in technology has got a great performance and costs impact on the overall result. It proves to be inevitable to optimise the logistics of paper document flow, to maximise the usage of technological elements, and to minimise the load on data concentrators and digital archiving space, as each repetition of processes and operations increases enormously the requirements for logistics, professional staff and digitisation costs.

Next works in the area of optimisation of mass digitisation processes will be led by:

1) Technological part: 2) Organisation and logistics, 3) Management and digital content organisation and reuse

Indicative list of Topic 1 objectives:

1. Optimisation of logistics and flows of printed material with respect to minimal movements between technological steps.
2. Optimisation of technological steps to reach the maximal usage of each technological element and minimise costs.
3. Optimisation of digitised data transfers between technological steps to achieve a better processing performance.
4. Introduction of further automated points of quality control with respect to minimise retries and achieve better quality near-to-actual-technology step and minimise unnecessary material movements.
5. Integration of all in-digitalisation-needed technologies into pluggable SOA/BPM based logistics and workflow system to maximise overall automation and minimise human intervention. Also to propose a reference model with industry SOA/ESB standards for inter-process communication and industry standard-based BPE/BPM workflow management system for use by any entity planning or carrying out mass digitisation projects.
6. Optimisation of acquisition, collecting, managing, long-term archiving of digital content, webharvesting, webarchiving.

The strategy development in Topic 1 is evaluation of the state-of-art and trends and new middle term strategy in mass digitisation in Slovakia. High level scientific awards and competitions consist in the implementation of: a) benchmarking of the world best available software and technologies b) accredited quality control testing and certified quality assurance c) implementation and validation of control methods of quality. The project benefits from expertise accumulated in the projects like IMPACT, Europeana, MINERVA etc.

6. Area of conservation and preservation (Topic 2)

A MSCOE researcher has knowledge of the newest technologies and best solutions in the field of mass conservation and digitisation, which will be applied in the project.

In the field of conservation and long-term preservation of analogue media, in 2012-2015 the SNL plans to implement within the DL&DA project the excellent results from the KNIHA SK *Project Preservation, Stabilisation and Conservation of Traditional Carriers of Information in the Slovak Republic* under the state basic research plan [5]. Besides solving technological and scientific issues, the benefit of the project was also in implementing the procedures of a scientific collaboratory [6].

The essential requirements for selecting a system and technology for conservation are contained within the criteria. In the public procurement process, the project will also take into account the *criteria and requirements for technology directly related to the conservation-based preparation of documents for digitisation.* Under the DL&DA project, the complete best available technology of Papersave Swiss (app. 10 mil. €) will be implemented by the SNL in 2013 under the DL&DA project, to be maintained at least until 2020. The SNL as a research partner of the KNIHA SK task possesses the results which can potentially improve the best available Papersave Swiss technology multiple times.

Objectives of Topic 2: Support of research in the field *of mass conservation and preservation (deacidification) sterilisation* through exchange of know-how and experience with partners; recruitment of experienced researchers; specialist dissemination and outreach to innovation capacity building activities. Specific subtopics for research and innovation are focused on the following two areas: 1) Implementation of industry SOA/ESB standards as described in Topic 1 for conservation and preservation technologies with producer's participation to achieve integration with industry standard BPM based workflows and logistics systems. 2) Sterilisation, conservation and digitisation – internal logistics optimisation with respect to achieve optimal performance and usage of these technologies. 3) Conservation – chemical process optimisation with respect to a short treat cycle.

The DL&DA project's goal is to: a) improve technological and technical equipment of laboratories, conservation and preparation worksites for professional treatment, conservation and restoration of items and special collections (library, archival, ...) in a close relationship with the information and communication infrastructure for acquiring, processing and protecting the content; b) implementation of the research results in the field of mass deacidification of ligno-cellulose-based information carriers in heritage institutions, which are directly connected with digitisation.

The potential MSCOE project main partners are Nitrochemie Wimmis AG (Papersave Swiss), the Slovak University of Tech-

nology, BelNovaman International Ltd., and Groupe Eurofins, the Swiss National Library. Transfer of know-how and technology from Nitrochemie and Slovak University of Technology will take place in accordance with the existing agreements (secrecy agreement and business contract between the SNL and Nitrochemie). The SNL also uses know-how within the KNIHA SK research consortium, which includes the Slovak University of Technology, Slovak Academy of Sciences (the Institute of Polymers), and the Slovak National Archives. Preservation of archival and library collections pertains to the main tasks of archives and libraries throughout the world.

Unfortunately, the SNL currently does not employ a new generation of experts in the field of mass conservation of written cultural heritage. This is an opportunity for the MSCOE project. Since 2006 the SNL has been building its Integrated Conservation and Digitisation Centre (ICDC) as its present-day organisational unit.

The ICDC concentrates the following activities: a) mass digitisation and digital archiving of cultural heritage materials; b) conservation (i.e. cleaning, restoration, mass sterilisation, deacidification, lyofilisation) of paper-based library and archival documents; b) research, development and education in the above areas especially under the auspices of the MSCOE.

It is, therefore, necessary to implement the MSCOE project and thus support the building of capacities in research, development and education and to set up the open cooperation in the EU.

Thanks to the effective link established between the academic sectors with cultural heritage sectors the unique approach was conceived which has enabled the development of the DL&DA national project. The uniqueness of the approach rests in mass industrial digitisation and mass industrial conservation *as one integrated technological and functional system.* Such approach is rare on both European and global scale.

7. Area of text analysis and knowledge mining (Topic 3)

The objectives of activities under this topic are: support of research in the field of text analysis and knowledge mining as a component of mass digitisation and digital content reuse through exchange of know-how and experience with partners; recruitment of experienced researchers; upgrading and acquisition of research equipment; specialist dissemination and outreach to innovation capacity building activities.

The outcome of this topic includes affordability, widespread availability of tools and services for releasing the economic potential of cultural heritage in a digital form and for adding value to cultural content in the educational, scientific and leisure context., and a wider range of users of cultural resources in diverse real and virtual contexts and considerably altered ways of experiencing culture in more personalised and adaptive interactive settings.

Upon completion of the national DL&DA project, the Slovak National Library will hold about 270 million pages of text in a digital form. Topic 3 goal is to implement new experiences and trends in working with extensive masses of digital texts. New world's trends show that the digitisation process is not the end of the process, but only its beginning.

Additionally, in presentation of the original content the information society requires options which emerge with the introduction of Web 2.0. For example, this concerns enriched content and supporting context, fostering multilingualism (which means that the original content is viewable and searchable through translation to other languages), and offering web links to already existing knowledge. This information is to be generated directly from the original text, it is not changed, but supplemented in combination with knowledge databases. These requirements also result from the Memorandum of Europeana member libraries as regards the methods of presenting digitised content.

Topic 3 is focused on supporting research using advanced methods of text analysis and natural language processing specialised in the Slovak language. The activities include pre-processing, classification, categorisation, clustering of text and knowledge extraction from source texts aimed at assisting the SNL in making decisions concerning text processing methods applied in mass digitisation processes (structural analysis, metadata assignment, tokenisation, lemmatisation, classification, clustering). There will be clear benefits of this interaction in the fact that currently, researchers concerned with linguistics, artificial intelligence, cybernetics, library and information science have not cooperated so far at a national level, and the national text digitisation project gives them a unique opportunity to apply theoretical knowledge into production practice to the users' benefit.

This publication is a result of implementing the "Memory of Slovakia: National Centre of Excellence in Research, Preservation and Accessibility of Cultural and Scientific Heritage" Project (ITMS:26220120061) supported by the Research & Development Operational Programme funded by the ERDF.

References

[1] *Digital Agenda for Europe: Digital Libraries Initiative. Europe's cultural and scientific riches at a click of a mouse.* 2010. Dostupne: http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm *Europeana* [website]: http://www.europeana.eu/portal/, Digital Agenda [website] http://ec.europa.eu/digital-agenda, Neelie Kroes [website] http://ec.europa.eu/commission_2010-2014/kroes/

[2] *Digital Agenda: Encouraging Digitisation of EU Culture to Help Boost Growth.* European Commission – Press Release. Brussels 28th October 2011. Available at: http://europa.eu/rapid/press-release_IP-11-1292_sk.htm?locale=en

[3] *Opinion of the European Economic and Social Committee on 'Unlocking the Potential of Cultural and Creative Industries* (Green Paper)' COM(2010) 183 final (2011/C 51/09) [viewed 2-12-2012] Available at : http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2011:051:0043:0049:SK:PDF

[4] KATUSCAK, D., DZIVAK, J., CAPKOVIC, M., HVOLKA, M., ARDO, O., GABRIELOVA, J.: *Digitalna kniznica a digitalny archiv* – narodny projekt [Digital Library and Digital Archives – National Project.] Operacny program informatizacie spolocnosti OPIS2. Implementacia 2010 – 2015. Martin : Slovenska narodna kniznica, Kompletný projekt so stavebnymi projektmi a dokumentaciou, ca 4000 p., 2011

[5] KATUSCAK, S., BAKOS, D., BUKOVSKY, V. et al: *Zachrana, stabilizacia a konzervovanie tradicnych nosicov informacii v Slovenskej republike (KNIHA SK)* [Rescue, Stabilisation and Conservation of Traditional Information Carriers in the Slovak Republic]. Kod statnej vyskumnej ulohy: 2003SP200280301. Slovenska technicka univerzita v Bratislave : Fakulta chemickej a potravinarskej technologie. C. spravy: 09/533. Archivne c. spravy: 09/533/2009. Druh spravy: Zaverecna sprava, 2003 – 2008). SIGNATURA: VS 185. Bratislava : STU, 2009. Sprava na zaverecnu oponenturu statnej vyskumnej ulohy. Project materials are available at: http://www.knihask.eu/

[6] KATUSCAKOVA, M., KATUSCAK, M.: *Recommendations for Scientific Collaboratories: Application of KM Findings to a Scientific Collaboratory.* Proc. of the 13th European Conference on Knowledge Management, vol.1, 2012. ISBN 978-1-908272-64-5.