



**I T H E A**



**International Journal**

**INFORMATION THEORIES  
&  
APPLICATIONS**



**2004 Volume 11 Number 3**



**International Journal  
INFORMATION THEORIES & APPLICATIONS**

ISSN 1310-0513

Volume 11 / 2004, Number 3

This issue contains papers from the International Seminar "Digitization of Cultural and Scientific Heritage"  
27 August – 3 September 2004, Bansko, Bulgaria

IJ ITA Editor in chief: **Krassimir Markov** (Bulgaria)

**IJ ITA International Editorial Staff**

Chairman: **Victor Gladun** (Ukraine)

<b>Adil Timofeev</b>	(Russia)	<b>Larissa Zainutdinova</b>	(Russia)
<b>Aleksey Voloshin</b>	(Ukraine)	<b>Laura Ciocoiu</b>	(Romania)
<b>Alexander Eremeev</b>	(Russia)	<b>Levon Aslanian</b>	(Armenia)
<b>Alexander Kleshchev</b>	(Russia)	<b>Luis F. de Mingo</b>	(Spain)
<b>Alexander Kuzemin</b>	(Ukraine)	<b>Martin P. Mintchev</b>	(Canada)
<b>Alexander Palagin</b>	(Ukraine)	<b>Milena Dobрева</b>	(Bulgaria)
<b>Alfredo Milani</b>	(Italy)	<b>Natalia Ivanova</b>	(Russia)
<b>Anatoliy Shevchenko</b>	(Ukraine)	<b>Neonila Vashchenko</b>	(Ukraine)
<b>Arkadij Zakrevskij</b>	(Belarus)	<b>Nikolay Zagorujko</b>	(Russia)
<b>Avram Eskenazi</b>	(Bulgaria)	<b>Petar Barnev</b>	(Bulgaria)
<b>Boicho Kokinov</b>	(Bulgaria)	<b>Peter Stanchev</b>	(Bulgaria)
<b>Constantine Gaidric</b>	(Moldavia)	<b>Plamen Mateev</b>	(Bulgaria)
<b>Eugenia Velikova-Bandova</b>	(Bulgaria)	<b>Radoslav Pavlov</b>	(Bulgaria)
<b>Frank Brown</b>	(USA)	<b>Rumyana Kirkova</b>	(Bulgaria)
<b>Galina Rybina</b>	(Russia)	<b>Stefan Dodunekov</b>	(Bulgaria)
<b>Georgi Gluhchev</b>	(Bulgaria)	<b>Tatyana Gavrilova</b>	(Russia)
<b>Ilija Mitov</b>	(Bulgaria)	<b>Valery Koval</b>	(Ukraine)
<b>Jan Vorachek</b>	(Finland)	<b>Vasil Sgurev</b>	(Bulgaria)
<b>Juan P. Castellanos</b>	(Spain)	<b>Vitaliy Lozovskiy</b>	(Ukraine)
<b>Koen Vanhoof</b>	(Belgium)	<b>Vladimir Jotsov</b>	(Bulgaria)
<b>Krassimira Ivanova</b>	(Bulgaria)	<b>Zinoviy Rabinovich</b>	(Ukraine)

**IJ ITA is official publisher of the scientific papers of the members of  
the Association of Developers and Users of Intellectualized Systems (ADUIS).**

IJ ITA welcomes scientific papers connected with any information theory or its application.

Original and non-standard ideas will be published with preferences.

*IJ ITA rules for preparing the manuscripts are compulsory.*

*The rules for the papers for IJ ITA as well as the subscription fees are given on [www.foibg.com/ijita](http://www.foibg.com/ijita).*

*The camera-ready copy of the paper should be received by e-mail: [foi@nlcv.net](mailto:foi@nlcv.net)*

Responsibility for papers published in IJ ITA belongs to authors.

General Sponsor of IJ ITA is the **Consortium FOI Bulgaria** ([www.foibg.com](http://www.foibg.com)).

**International Journal "INFORMATION THEORIES & APPLICATIONS" Vol.11, Number 3, 2004**

**Printed in Bulgaria**

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria,  
in collaboration with the V.M.Glushkov Institute of Cybernetics of NAS, Ukraine, and  
the Institute of Mathematics and Informatics, BAS, Bulgaria.

Publisher: FOI-COMMERCE - Sofia, 1000, P.O.B. 775, Bulgaria. [www.foibg.com](http://www.foibg.com), e-mail: [foi@nlcv.net](mailto:foi@nlcv.net)

© "Information Theories and Applications" is a trademark of Krassimir Markov

**Copyright © 2004 FOI-COMMERCE, Publisher**

**Copyright © 2004 For all authors in the issue.**

All rights reserved.

ISSN 1310-0513

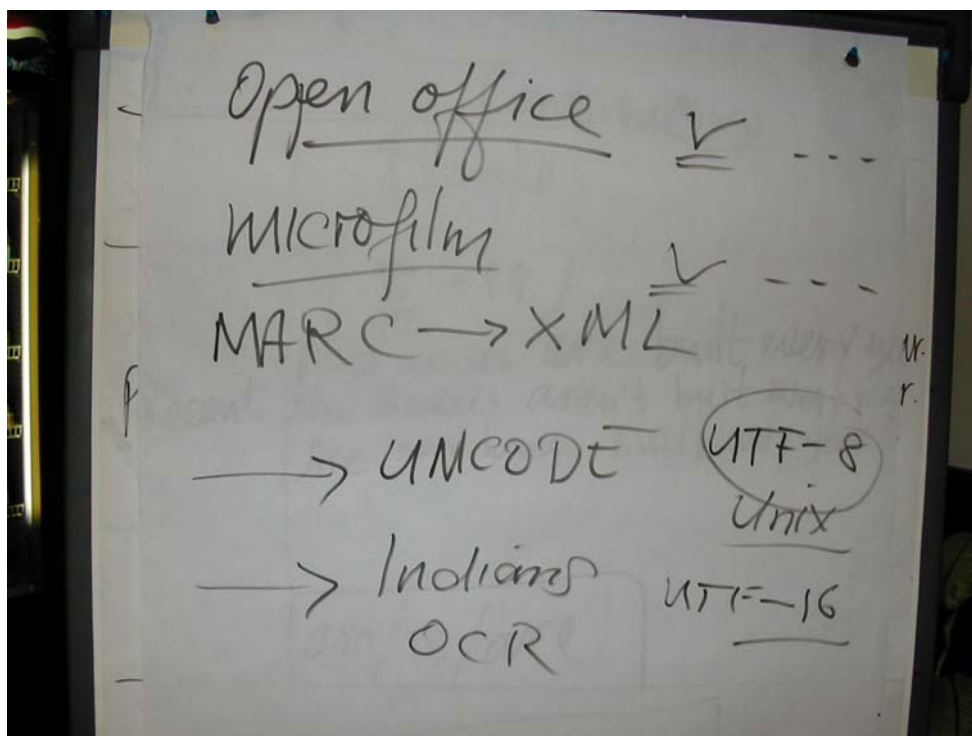
## THE MANY FACES OF DIGITISATION

**Milena Dobрева**

This special issue of the International Journal "Information Theories and applications" presents papers delivered during the international seminar "Digitisation of cultural and scientific heritage". The seminar was held in Bansko, Bulgaria from 27 August to 3 September 2004. It accompanied the kick-off meeting of the project *Knowledge Transfer for the Digitisation of Cultural and Scientific Heritage in Bulgaria* (KT-DigiCULT-BG), supported by the Marie Curie programme, Framework Programme 6 of the European Commission.

The event attracted some 50 participants from 15 countries: Bulgaria, Czech Republic, Germany, Denmark, France, FYROM, Greece, Ireland, Italy, Malta, the Netherlands, Serbia and Montenegro, Turkey, UK, and USA. Participants included representatives of the coordinator, the Institute of Mathematics and Informatics, (Bulgaria) and four project partners: Det Arnamagnæanske Institut (Københavns Universitet, Denmark), Trinity College (Dublin, Ireland), Charles University (Prague, Czech Republic), and the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demkoritos" (Athens, Greece). In addition, participants from a number of European institutions, which already have impressive digitisation activities in the last years, presented various issues related to digitisation and thus contributed to the successful exchange of experiences. Last but not least, a number of presentations outlined the existing experience in Bulgarian institutions from the cultural and scientific heritage sector and from organisations involved in research on the topic.

The project is presented in the first article of this compendium and we will not give here more details on it. Unlike the traditional prefaces, I will not present what follows – the actual topics are much more diverse than ones written on the board and digitally captured during one of the discussions...



A distinguishing feature of the meeting was the variety of topics, approaches and results presented. I hope that this variety, which was found very stimulating by the participants, will be also inspiring for the readers of the volume.

The papers do not deviate from the original style of expression of their authors.



Shortly after the meeting was held, the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences created a Department on Digitisation of Scientific Heritage (<http://www.math.bas.bg/digi/index.html>).

Further scientific events planned by the department include (for more information please visit the website of the department):

- an international workshop on **Annotation for Written Cultural Heritage Resources** which will be held in Prague, 6-13 March 2005;
- an international workshop on **Computational Tools for the Librarian and Philological work in Cultural Institutions**, which will be held within the frameworks of the 34<sup>th</sup> Spring conference of the Union of Bulgarian Mathematicians, 6-9 April 2005 in Borovets, Bulgaria;
- the 10<sup>th</sup> EIPub conference (International Conference on Electronic Publishing) which will be held in Sofia in June 2006 under the motto 'The Digital Spectrum: from Technology to Culture'.

Members of the department also actively contribute to the organisation of The First South-Eastern European Digitization Initiative (SEEDI) Conference **Digital (Re-)Discovery of Culture (Physicality of Soul) – Playing. Digital** (<http://www.ncd.matf.bg.ac.yu/seedi/events/seediConference.html>), to be held on 11–14 September 2005 in Ohrid, FYROM.

This volume is the first publication of the Digitisation of Scientific Heritage department at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences. As a head of the Department, I would like to express my hope that its work will be as diverse and rich as this collection of papers.

---

#### Author information

**Milena Dobreva** - Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev St., bl.8, Sofia-1113, Bulgaria; e-mail: [dobreva@math.bas.bg](mailto:dobreva@math.bas.bg)

---

## DIGITAL PRESERVATION AND ACCESS TO CULTURAL AND SCIENTIFIC HERITAGE: PRESENTATION OF THE KT-DIGICULT-BG PROJECT

Milena Dobreva and Nikola Ikonov

*Abstract:* The fast development and wide application of digital methods, combined with broadened access to the Internet and falling computing costs, have created intense interest in electronic presentation and access to cultural and scientific heritage resources. Information technologies have offered cultural institutions new opportunities for the presentation of their holdings, which are now made accessible not only to the specialists, but also to the citizens and interested parties worldwide.

The paper presents an overview of the Bulgarian experience in the field of digital preservation and access and on-going work on the project "Knowledge Transfer for the Digitisation of Scientific and Cultural Heritage to Bulgaria" (MTKD-CT-2004-509754) supported by the Marie Curie programme of the FP6 of the EC.

*Keywords:* digitisation, cultural and scientific heritage.

---

### Introduction

The field of digitisation of cultural and scientific heritage is within the priority areas for the European Union as an inevitable part of the understanding of our common grounds and local similarities and diversities. While in the previous decades the exposure and discussion over scientific and cultural heritage were a privilege to small scientific communities, now this heritage reveals greater interest and can be exposed easily to the general public with the help of modern information technologies.

This is an important development stimulus. However, the practical work in this field requires substantial efforts and much specialised expertise both to prepare the resources in electronic form and to present them properly to various audiences. The project presented in this paper aims to remove currently existing research and practical work gap for Bulgaria. It in fact serves as a pilot effort for a country within the Balkan region, and has all chances to multiply the transfer knowledge via know-how transfer to other countries with similar needs.

We witness for over one decade now that while the interest to methods and tools for presentation of cultural heritage in Bulgaria grows, the real efforts usually end up with small demonstration projects. To move the current state from this point, we applied to the Marie Curie programme (action Transfer of Knowledge), which enables us to bring on a regular basis experienced researchers to the Bulgarian host (Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences). The design of the programme should contribute to the development of synergies between various institutions, both on national and international scope. The ratio of experienced and more experienced researchers is 1:1 and this actually should help to avoid the generation gaps which cover the field in an unbalanced way concerning experience and practical knowledge.

Even in the countries from the European Union, the expertise of specialists working in the field of digitisation is gained mostly by individual practice. When a young researcher has the chance to become a part of an experienced group, he/she would gain necessary skills to work on a good quality. Otherwise, many institutions establish small-scale projects learning basically from pitfalls. This tendency is not a positive one, because it results in the production of scattered resources which could not be interconnected.

---

### Overview of the Bulgarian Experience

#### *Collections in Bulgaria*

In Bulgarian repositories over 12,500 manuscripts of Slavonic, Greek, Latin, Islamic and other origin are preserved. Bulgarian institutions also keep the third largest collection in the world (following Italy and Greece) of

epigraphic inscriptions from Antiquity. It is obvious that such materials are of interest not only for the local community but also on a European and global scale.

This set of resources is still hardly accessible in its fullness not only to foreign experts but also to regional experts. Electronic cataloguing and digital preservation are still not popular in the region.

#### *Brief History of Digitisation Initiatives*

The first initiatives in this field were launched about 15 years ago by research institutions and companies. A national strategy and funding for digitisation programmes has not been available since the very beginning of works. Actual activities on digitisation have not been done on large-scale basis yet.

The first field where computers were implemented in the late 80s and early 90s of the last century was *cataloguing*. The **ISIS** library cataloguing software was introduced in Bulgaria and tested in the National Library "St. Cyril and St. Methodius". However, major cataloguing effort has not been started in that time, because of the limitations of the model. Already in this first introduction of computers the expectations that the new technologies would assist the work of specialists in medieval manuscripts, had not overcome the fact that these specialists did not like the model built in the system.

A next effort in the field of electronic presentation of manuscript data was the project called **The Repertorium of Old Bulgarian Literature and Letters**. It started as "an archival repository capable of encoding and preserving in SGML (and, subsequently, XML) format of archeographic, palæographic, codicological, textological, and literary-historical data concerning original and translated medieval texts represented in Balkan Cyrillic manuscripts" [Repertorium]. The project grew out of an initiative of D. Birnbaum (University of Pittsburgh), A. Bojadžiev (University of Sofia), M. Dobreva (Institute of Mathematics and Informatics, Bulgarian Academy of Sciences), and A. Miltenova (Institute of Literature, Bulgarian Academy of Sciences) in 1994. The computer model is discussed in [Dobreva 00]. Currently there are 300 manuscript descriptions available in the Institute of Literature of the Bulgarian Academy of Sciences. The basic publication on this endeavour is [MB 00].

**Adoption of MASTER-Manuscript Access through Standards for Electronic Records for Cataloguing of Old Bulgarian Manuscripts.** MASTER was a European project funded under the Framework IV Telematics for Libraries programme [MASTER]. It developed a TEI<sup>1</sup>-conformant DTD<sup>2</sup> for medieval manuscripts with the ambition to serve the needs of all repositories in Europe, and software for making and visualising records on manuscripts. The MASTER standard (with some small revisions) was adopted by the TEI in 2003. The National Library "St. Cyril and St. Methodius" and the Institute of Mathematics and Informatics joined MASTER as associated members. They made 30 descriptions of manuscripts with two basic aims: to test descriptions providing data both in English and Bulgarian, and to supply examples coming from the non-Latin literary tradition.

Although the *Repertorium* and *Master* project are oriented to the same standard framework, the models underlying both projects are not similar (i.e. they contain different sets of elements structured in different ways).

In addition to these cataloguing efforts, some companies created CD-ROMs (four CD-ROMs already exist, two of manuscripts from the National Library "St. Cyril and St. Methodius", one of Macedonian coins and one of Bulgarian Iconography).

Beyond the manuscript field, an interesting project was launched to present one of the most interesting historic buildings in Bulgaria – the Boyana Church in 2000-2002 [Boyana]. This is the first computer-based 3D model of a Bulgarian monument of culture. Author of this project is Trifon A. Trifonov.

#### *Experience of IMI-BAS*

Specialists from IMI-BAS took part in the Repertorium and MASTER project mentioned above thus gaining expertise in presentation of **medieval written heritage**.

---

<sup>1</sup> Text Encoding Initiative, see [TEI].

<sup>2</sup> Document Type Definition

In addition, one of the obvious interests of this community is related to **digitisation of mathematical heritage**. The Institute of Mathematics and Informatics has been hosting a growing group of people accomplishing various related activities with mathematical texts. These activities include the whole preprint process of the key Bulgarian mathematical journals:

- Serdica Mathematical Journal Pliska Studia Mathematica Bulgarica Proceedings of the Annual Conference of the Union of the Bulgarian Mathematicians and also of the interdisciplinary review
- Comptes rendus de l'Académie Bulgare des Sciences In collaboration with Lefkowitz & Co.

The group took part in the digitisation of the *Jahrbuch Für Mathematic*. Other fields of recent work of specialists from IMI include **study of chronological distribution of historical artefacts, edutainment, encoding of Early Cyrillic and Glagolitic alphabets, presentation of immovable heritage**.

In addition, IMI hosted or contributed to the organisation of a number of international events related to digitisation of cultural heritage. This places it in a very important position for **disseminating and raising awareness activities**. Here we should list The First International *Conference Computer Processing of Medieval Slavonic Manuscripts* was held in Blagoevgrad in 1995 (see [BBDM 95]). The UNESCO workshop *Text Variety in the Witnesses of Medieval Texts* was held in Sofia in 1997 (see [Dobrev 98]). The Institute of Mathematics and Informatics organised a series of summer schools in 1998, 1999 and 2002 aimed at enlarging the international community of young specialists who would share their experience, and form co-operation ties, as well as a workshop on Digitisation of Cultural Heritage within the frameworks of the 1<sup>st</sup> International Congress of the Mathematical Society of South-Eastern Europe [NCD 2004].

#### *Some Conclusions*

From this brief presentation of digitisation-related activities, it is obvious that the major field where experience was gained is electronic cataloguing of manuscripts. Yet, we cannot speak about a complete consensus between different teams, using separate catalogue descriptions. This fact causes diversity in approaches, but it also illustrates the lack of efforts for integration, which on the long term run leads to dissolved results. Such disagreement is also contradictory to the European recommendations expressed in the Lund principles.

---

### **The KT-DigiCult-BG Project**

---

The fast development and wide application of digital methods, combined with broadened access to the Internet and falling computing costs, have created intense interest in electronic presentation and access to cultural and scientific heritage resources: original manuscripts, early printed books, epigraphic inscriptions, etc. Information technologies (IT) have offered cultural institutions new opportunities for the electronic presentation of their holdings, which are now made accessible not only to the specialists, but also to the citizens and interested parties worldwide.

On this setting, the electronic resources available for the Slavonic countries, first of which became members of the European Union in 2004, are still scarce. In a small country like Bulgaria it is impossible, due to the lack of specialists, and economically not efficient to form digitisation groups attached to the various cultural and scientific heritage institutions. The work on this project will strengthen the experience gained by the Institute of Mathematics and Informatics in the digitisation field and develop it further through knowledge acquisition and transfer measures. Thus the Institute will develop as a national centre of best practice in the field, and will be able to support on-going initiatives in the digitisation sphere.

The basic activities envisaged in the project will contribute to change the current state in the digitisation field in Bulgaria through:

1. Designing an integrated approach for the presentation of the material, which is large in volume, rich in language variation and multimodal as a computer presentation;
2. Implementing IT framework suitable for appropriate presentation of the local cultural heritage within the European electronic space;

3. Establishing the bases for a cost-effective and fast semi-automatic and automatic content-sensitive annotation of the word mass of the written cultural sources;
4. Integrating the experience of EC partners and accession countries, thus decreasing the gap between the state-of-the-art and real work in Bulgaria and the rest of Europe.

Digitisation field is combining knowledge from several different specialised fields (the project considers digitisation in the broad sense, including presentation of a variety of data on a cultural artefact in a computer form, not just digital imaging, i.e. texts, structured data, audio, etc.). To fulfil project goals, the host will cooperate with partners who had gained expertise in different fields. This will help to build a well-balanced team within the host which is not concentrating on one of the problems in the field, but is able to approach it in creative way, taking into account the methods and techniques applied for digital image processing, digital document archiving and cataloguing, information retrieval, distributed systems, classical and historical lexicography, encoding and document type definitions.

Providing digital access to cultural heritage has an important influence on the actual preservation of the originals. To these economic and societal impacts of digitisation we could add the effect of improved visibility of cultural artefacts for the citizens. The profound comprehension of the current settings and historical reasons for the present status quo lays in the better understanding of where our sources are stated in the Lund principles of 2001.

Our project fits most to the following major trends envisaged in the Lund action plan (ordered according to their importance):

- Action 4b: *Sustainable access to content* – it will be assured through the framework offered.
- Action 3a: *Good practice examples and guidelines* – they are per se incorporated into KT-DigiCult-BG and disseminated widely through the manuals which will be created.
- Action 3b: *Competence centres* – the host organisation from Bulgaria will have the real opportunity to grow as such centres which will 'spread the word' further. In addition, the close ties with partners will provide to them valuable feedback and probably would boost their own development as competence centres.
- Action 1d: *Supporting coordination activities* – the project defines a stable inter-cooperation, a core of a network, which did not exist in the past in this field.

Through the KT-DigiCult-BG project we intend to overcome well-known barriers present in the associated countries community and identified in the Lund Principles:

- *Fragmentation of approach* – this is what we have been witnessing in the last decade in Bulgaria, and the project will contribute to change this tendency;
- *Obsolescence* – we are targeting at learning from the best experience in the EC, which guarantees the state-of-the-art in this transfer of knowledge project;
- *Lack of simple, common forms of access for the citizen* – this is currently a fact for the Slavonic heritage, and also for the Bulgarian environment;
- *Data Protection and Intellectual property rights* – they are recognised in our effort and could serve as a real life example in the future endeavours in this field.

Our strongest contribution is in the following areas endorsed by the European experts in the Lund Principles:

- *An accessible and sustainable heritage* – through developing state-of-the-art framework and tools which would place in the e-European space the mediæval Slavonic written heritage which is now available as small-size and scattered resources;
- *Support for cultural diversity, education and content industries* – by exposing a cultural heritage, which is now missing in the electronic space;
- *Digitised resources of great variety and richness* – adding to the existing resources one more significant group.



---

This project attracts as participants key organisations from EC member countries (Charles University Prague, Trinity College Dublin, Copenhagen University, Institute of Informatics and Telecommunications at NCSR Demokritos Institute in Athens). Some of them already cooperated, basically in training activities, which contributed to raise the awareness on the importance of the digitisation of cultural heritage in Bulgaria. Therefore, KT-DigiCult-BG improves the maturity and the high quality approach in the field. The previous experience contributes to transfer of knowledge of highest possible quality. Up till now, specialists in Bulgaria mastered in this field if they had the chance to work with a leading specialist abroad, or devised their knowledge from own experience. The project gives a chance in future to young people to receive structured and well-balanced theoretical and practical framework and will boost real work in the field. The development of local centres of such high quality in Bulgaria is one of the measures to prevent brain-drain.

Basic fields of work which will be supported through visits of incoming researchers include but are not limited to:

- General methodology and practical setting for digitisation of cultural and scientific heritage.
- Digitisation of medieval manuscripts (incl. digital imaging, cataloguing, text representation, electronic publishing).
- Digitisation of mathematical texts and building digital mathematical library of works of Bulgarian mathematicians.
- Virtual reality applications for presentation of immovable cultural heritage.
- Audio archives: methods for digitisation and restoration.
- Application of quantitative methods for the study of data related to the cultural heritage.
- Applications of edutainment to cultural heritage studies.

During the first project year, incoming researchers included Dr. Matthew Driscoll from Copenhagen University who worked together with project team members on an XML editor for cataloguing mediaeval Bulgarian manuscripts; Boris Shishkov from Delft University of Technology, the Netherlands, who will develop an electronic brokerage system for sites presenting cultural and scientific heritage, and Philip Zrantchev from the University of Reading, UK, who develops an Old Cyrillic UNICODE font based on Codex Suprasliensis script.

Thus our project already contributes to improve the quality and contents of future national and international work in digitisation of cultural and scientific heritage by very intensive and well-designed transfer of knowledge, which not only brings knowledge to Bulgaria, but is aimed at its best integration according to the local needs. The project will have important impact on the future developments of electronic presentation/publishing/preservation of cultural heritage in Bulgaria, but also will serve as an example for future work in countries with similar economic and cultural settings.

---

## **Conclusion**

---

In the period between the submission of the project and its start the priorities in the field of digital preservation of and access to cultural and scientific heritage resources of the Information Society Technologies thematic area of FP6 changed. Dealing with complex and dynamic objects and new knowledge technologies, visualisation and virtual reality are now in the focus of EC support.

On this setting, Bulgaria still has to solve numerous problems related to making available cultural and scientific heritage resources in digital form. We hope that one of the feasible outcomes of this project will be to produce resources in digital form at least in the field of mathematical heritage and archival collections.

## Bibliography

- [Boyana] <http://www.boyanachurch.org/> – website of the Boyana Church
- [BBDM 95] Birnbaum, D., A. Bojadjiev, M. Dobрева, A. Miltenova, ed. *Computer Processing of Medieval Slavic Manuscripts. Proceedings of the First International Conference*. 24–28 July 1995. Blagoevgrad, Bulgaria. Sofia: Professor Marin Drinov Academic Publishing House. 1995. ISBN 954-430-417-7.
- [Dobрева 98] Dobрева, M. (Ed). *Text Variety in the Witnesses of Medieval Texts. Proceedings of the International Workshop*. Institute of Mathematics and Informatics. Sofia, 21–23 September, 1997. Sofia: Institute of Mathematics and Informatics. 1998. ISBN 954-9650-02-2.
- [Dobрева 00] M. Dobрева, *A Repertory of the Old Bulgarian Literature: Problems Concerning the Design and Use of a Computer Supported Model*, In: A. Miltenova, D. Birnbaum (eds.), *Medieval Slavic Manuscripts and SGML: Problems and Perspectives*, Sofia, Academic Publishing House, 2000, pp. 91-98.
- [Lund Principles 01] [http://www.cordis.lu/ist/ka3/digicult/lund\\_principles.htm](http://www.cordis.lu/ist/ka3/digicult/lund_principles.htm) – *eEurope: creating cooperation for digitisation (Lund Principles)*
- [MASTER] <http://www.cta.dmu.ac.uk/projects/master/>, website of the MASTER project.
- [MB 00] A. Miltenova, D. Birnbaum (eds.), *Medieval Slavic Manuscripts and SGML: Problems and Perspectives*, Sofia, Academic Publishing House, 2000, 372 pp.
- [NCD 2004] Review of the National Center for Digitisation, No 4(2004), Serbia and Montenegro, a special issue with papers presented at the minisymposium Digitisation of Cultural Heritage, Borovets, 2003.
- [Ognjanović 02] *National Center for Digitization*, In: *Review of the National Center for Digitization*, 1/2002.
- [Repertorium] <http://clover.slavic.pitt.edu/~repertorium/index.html> – website of the *Repertorium of Old Bulgarian Literature and Letters*
- [Ross et al. 03] S. Ross, M. Donnelly, M. Dobрева, *New Technologies for the Cultural and Scientific Heritage Sector (DigiCULT, Technology Watch Report 1)*, European Commission, 2003, 196 pp. ISBN 92-894-5275-7.
- [TEI] <http://www.tei-c.org/> – *Text Encoding Initiative Website*
- [Tutorial] <http://www.library.cornell.edu/preservation/tutorial/> – *Moving Theory into Practice: Digital Imaging Tutorial*

### Web resources

- <http://palimpsest.stanford.edu/> – CoOL, Conservation OnLine, incl. <http://palimpsest.stanford.edu/bytopic/imaging/> – Digital Imaging
- [http://www.bl.uk/gabriel/services/lists\\_generated/services\\_digital\\_en.html](http://www.bl.uk/gabriel/services/lists_generated/services_digital_en.html) – Gabriel (links to collections of Europe's national libraries that have been digitised)
- <http://www.hatii.arts.gla.ac.uk/SumProg/DigiSS03/urls.htm#DigiHerAssets> – The Humanities Advanced Technology and Information Institute, University of Glasgow, Links to digitisation resources and sites
- <http://www.isos.dcu.ie> – Irish Script on Screen (ISOS)
- [http://www.kb.nl/kb/resources/frameset\\_kb.html?/kb/sbo/digi/verhanen.html](http://www.kb.nl/kb/resources/frameset_kb.html?/kb/sbo/digi/verhanen.html) – Research and development of electronic access to Medieval Illuminated Manuscripts from the Koninklijke Bibliotheek, The Netherlands
- <http://www.memss.arts.gla.ac.uk/> – The Digitisation of Middle English Manuscripts

## Authors' Information

- Milena Dobрева** – Chair of Dept. on Digitisation of Scientific Heritage, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev St., bl. 8, Sofia-1113, Bulgaria, e-mail: [dobрева@math.bas.bg](mailto:dobрева@math.bas.bg)
- Nikola Ikonov** – Chair of Laboratory on Phonetics and Speech Communication, Institute for Bulgarian Language, BAS, Shipchenski prohod 52, Sofia-1113, Bulgaria, e-mail: [nikonov@ibl.bas.bg](mailto:nikonov@ibl.bas.bg).

## THE EXPERIENCE AT TRINITY COLLEGE DUBLIN

Mícheál Mac an Airchinnigh

*Abstract: This is a script of a play(ing). A performance once enacted and then reflected upon is herein described.*

*Keywords: bridging, culture, image, playing, togetherness, physicality, picture, re-creation, re-discovery, soul*

*"Play, like imagination, could mend the broken soul." [Kane 2004, 46]*

---

### Prologue

This paper is the script of the short fifteen minute play(ing) performed in the opening of the KT-DigiCult-BG International Seminar in Bansko, Bulgaria. Like any play(ing), it consists of scenes grouped into acts. Being a script, it is naked. In particular, its nakedness is due to the absence of pictures and images which are an integral part of both the play(ing) and this paper. Fortunately, images of the first play(ing) are available on the KT-DigiCult-BG World-Wide Web (WWW) site. Googling reveals all. Each picture or image mentioned in this paper is indexed by the slide of the presentation for ease of reference. The original title of the presentation was The Experience of Trinity College Dublin.

The rest of the paper is organized as follows.

1. **Setting the Scene of the Play(ing).** In the first act of any play(ing) one must set the scene. How is one to do that when it seems that one is not in a theatre and the audience does not expect that they are about to see a play(ing)? My solution was then, and here now in this paper, to be poetical – the paper begins with a poem, in three languages: Bulgarian, Irish Gaelic, and English, in that order. The purpose is to disorganize rational thought and to appeal directly to the soul (through sound impressions ... an experience impossible to re-create in the written literature) and images (also not possible here due to page and printing limitation).
2. **Irish Gaelic Cultural Heritage.** A brief account of some conventional digitization of cultural heritage is presented in order to justify formally the presence of the playing performance at the KT-DigiCult-BG International Seminar.
3. **Bridging** is first of the trinity of key ideas of the paper. The other two are playing and togetherness. It is noted here that the ideas are processes, practical actions. I take the position that "to see", i.e., to experience practically one's own culture through the eyes of the other, especially an other from a different culture, is the most rewarding experience possible and especially in the context of [ DrDC ]. Bridging is the process of the mutual cultural enrichment of the couple, and by extension, of the many.
4. **Playing.** "The rhetoric of play as the imaginary [...] idealizes the imagination, flexibility, and creativity of the [...] human play worlds." [Sutton-Smith, 1997, 11]. Culture emerges from play. It is not so easy to see how one's culture emerges from playing within the culture of an other over a short period of time. The [ DrDC ] game suggested, and very vaguely outlined in the paper is intended, to explore this idea in practice.
5. **Togetherness** is a neologism which means "to gather together" ideas and people especially over the WWW.
5. **Epilogue.** Why Me ? Будител ? This is an "apologia" in the classical sense of a "defense".

---

### Setting the Scene of the Play(ing)

The theme is the Digital re-Discovery/re-Creation of Culture. The first part "Digital re-Discovery of Culture" is taken from the title of an EduTainMent paper [Sotirova 2004b]. This paper and its earlier companion [Sotirova 2004a] informs and inspires much of what is here presented. I can explain the essence of the philosophy underpinning [ DrDC ], an abbreviation of Digital re-Discovery/re-Creation of Culture, by a simple analogy.

In the early part of the 1970's I organized a cycling tour for young students (aged approximately 16) to Germany, a country which I had not yet visited. The cyclists were to stay in Jugendherberge (Youth Hostels), a different one each night. The routes had to be planned and timed, taking into account departure from and return to Ireland. At that time, one needed detailed maps (suitable for cyclists), a list of Jugendherberge, a telephone, and a practical knowledge of the German language. All these things were acquired and put to good and successful use. I remembered stopping one day in a small German town along the Rhein, and experiencing my first taste of a typical Turkish sweet pastry at lunchtime. Also at the same time I gave an interview in my best (broken) German to the local newspaper, a copy of which (together with photograph) was sent to me later in Ireland. This typifies for me today, in some way, what it means for **physicality of soul**. Specifically I mean this : through the tools of map, e-communication (telephone) and language (German) I had set up a "discovery of culture" for myself and my students. But the actual experience of the culture did not take place until I and all the others were physically present in place. Then the soul-anticipation was physically realized. With this analogy I will now try to explain the script of the play(ing).

The target audience and original setting of the play(ing) was Bulgarian, an audience with a good knowledge of the English language. In the context of multi-culturalism, it is the norm that the deepest and richest culture of the (majority of the) audience be addressed first. Therefore the Bulgarian theatre audience must be addressed first. Others must wait. Hence in the scene setting the first soul provocation is a poem in Bulgarian.



## БЕЗ ИМЕ

— — — — —

Капки дъга  
падат от очите ми.  
Виждам бяло  
в болка и тъга.  
Други виждат без цвят,  
безцветно бяло —  
цвет няма.  
Моята душа вижда в цвят.

(С) Михаел Мак ан Архинг (поез), юли 2004  
(С) Калина Сотирова (преводач), юли 2004

— — — — —






Slide No.2 [of 11]

Let us **imagine** that there is a picture [slide 2] of the newly opened Mostar Bridge [Bridge\_Mostar, 2004] facing this poem in Bulgarian. Imagine also that there is no caption on the picture. The picture clearly shows a bridge but "without a name", just like the title of the Bulgarian poem БЕЗ ИМЕ. The picture is intended to point to the poem in all sorts of ways. Later we will see that the idea pointed at by both poem and picture is the concept of **bridging**. The bridging concept is also reinforced in the Bulgarian by the use of the word дъга (rainbow). There is a certain direct correspondence between the arc(h) of the дъга and the arc(h) of the Mostar bridge. But the title of the poem "without a name" is loaded philosophically. By this I mean that once we use words to name some reality then we assume that we know everything about that reality. Every such naming word is culturally loaded. But there are naming words which may block understanding and even generate hostility. In my analogy to explain [ DrDC ] I deliberately mentioned a real life experience, using names German and Turkish, in the same context. At this time of writing it has been announced by the European Union that Turkey may apply for accession. Such an announcement did not meet with universal approval, especially within Germany (and France and others).

*Poet's commentary:* I am only a beginner (1 year) in Bulgarian language. But poetically, I notice sounds of “a” in дъга, тъга, няма, and of “o” in бяло, twice. The rhythm of the poem is pleasingly stopped twice with the consonant of “ят”.

With a view to multi-culturalism which was more than Bulgarian and English in the setting of the original play(ing), the author/poet felt a strong need to exhibit his Irish Gaelic culture. The next poem read out in Gaelic served two purposes. First, the author/poet was the only one to understand (of this I was 99% certain). Few, reading this Gaelic text given here, will understand it now. But, it is important to know that a) Irish Gaelic is said to be the most ancient spoken and written “European” language, and b) its culture and the re-Discovery/re-Creation of such has been, and still is, a matter of cultural life and death among Irish people both at home and abroad. Hence all in the audience are now provoked by a certain strangeness. There are *two* ways, of all the many possible ones, in which I **imagine** that the audience might engage: to **listen** to the sound of the poem and to **look** at the picture facing it, about which I will speak below. One might also imagine that the audience might focus on the performer himself as the “centre of attraction” at this point. Such would be the unspoken totality of meaning of any solo performer on stage.

The second purpose in presenting the Gaelic poem was to heighten the tension in the performance and to raise some fundamental questions. Is this poem the same as the previous poem? In other words, is it a translation? Then, one might further ask which came first?



## GAN AINM

Braonta thuar ceatha  
a thiteann ó mo shúile.  
Chím bán  
i bpian is i gcrá.  
Chíonn an mhuintir eile soiléir –  
bán soiléir.  
Dath de dhí.  
Chíonn m'anam dath.

(C) Mícheál Mac an Aircinnigh (Fíle), Iúil 2004  
(C) Mícheál Mac an Aircinnigh (Aistritheoir), Iúil 2004





Slide No.3 [of 11]

Now let us **imagine** that there is a picture [slide 3] of a single stemmed white rose (бяла роза) facing this poem. The white rose is set against a background of green colours with a small splash of red in the top right corner. What might the picture point to in this case? What is the relation between the picture facing this poem and the picture facing the previous poem? Is there a common connection?

Now is the time to reveal some of the inner connections. Colour is a key component of the play(ing). Colour is exhibited in pictures and in words. The colour white is very special. The splitting of white light by a prism gives rise to the seven colours of a rainbow (дъга). Therefore, I am using white as colour to bridge the two poems. But only I the author/poet knew this at the time. Hence, the tension raised in the audience must soon be resolved.

*Poet's commentary:* The Irish rhyming vowel-sounds I used are naturally very different from the Bulgarian. Noteworthy is the use of “chí”, the (northern Irish) verb “to see”, occurring at the beginning of 3rd, 5th, and 8th lines. There is a further subtlety only apparent to one who understands Gaelic. There are two key words used in the poem, “ainm” (name) in the title and “anam” (soul) in the last line. They sound very alike and bracket the entire poem in a pleasing way.

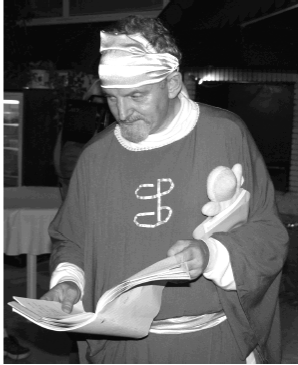
The tension brought to a height by the Irish Gaelic poem must now be completely resolved for all by the reading of the poem for the third time in the *lingua franca* of the audience — English.




## NO NAME

Rainbow drops  
 fall from my eyes.  
 I see white  
 in pain and sorrow.  
 Others see plain,  
 plain white.  
 No colour.  
 My soul sees colour.

(C) Micheál Mac an Airchinnigh, July 2004.





Slide No.4 [of 11]

Let us **imagine** a picture [slide 4] facing this poem, a picture of the author/poet dressed up in theatrical costume. It is very colourful and this colourfulness is the final arc(h) of the bridging set up by the three poems as pillars. The picture will also be an arc(h) to another one at the end of the play(ing) giving a unity to the entire thesis to be put forward and highlighting the significance to be paid to the concept of **bridging**.

Now that something of the staging has been explained, I want to point out that the sequence of poetic arc(he)s Bulgarian → Irish → English is deliberate. The poem was written in English first, translated by Kalina Sotirova into Bulgarian and finally by the author from both Bulgarian and English into Irish. The moral is: "Order of presentation need not be order of creation." Hence: "Order of re-discovery of culture need not be order of re-creation of culture." Finally, it is important to note that the real reason for giving three language versions of a single poem is to point out a universal reality behind culture and language. Each translator (of language pairs) will be able to bring the image of the poem into her/his culture. This is what we want to do in [ **DrDC** ] multi-culturalism.

*Poet's commentary:* notice the sound "ain" in **rainbow**, **pain**, **plain**. This was *not* planned. I mention this *a posteriori* analysis of my own poem in order to show that it is often the case that the poetic music arises quite unexpectedly. Translations of a poem from one language to another are always very difficult. Since the poem was written first in English, it surprises even me that there should be two very different kinds of poetical music sounding in two other very different languages. Finally, that there is a deep scientific and **physical** meaning behind the poem(s), I leave as a challenge to the reader. The only hint I give is this. Is there is a unique position, in which one must be, to see white turned into rainbow light? How can one have rainbow drops from tears of eyes? Where is the one who sees? Where is the light? Finally, that such translations of a single poem might be possible is a practical example of (cultural) **bridging** upon which we want to focus later.

---

### Irish Gaelic Cultural Heritage

---

The original title of this work and paper arising out of the KT-DigiCult-BG International Seminar, Bansko, Bulgaria, 2004, was the **Experience of Trinity College Dublin** (TCD) <<http://www.tcd.ie/>>. What was this experience? From the University of Dublin's point of view, the primacy of experience dealt with the Book of Kells (written c. 800 CE). The author could only report second-hand that such experience existed. That there might be other experiences, the author could only report second-hand also. But before all these second-hand experiences, the author engaged in the *Representation and Reconstruction of Chester Beatty Papyri* [Mac an Airchinnigh, 1991], which inspired to a certain extent the enthusiasm of Milena Dobrova to work on Medieval Slavonic Manuscripts [Dobrova, 1994, 21—22].

Since the author's experience transcended that of the Institution (at least historically) then much of the experience recounted in this paper is that of the author's. Nevertheless, it was important in the international and multi-cultural context to note that TCD had engaged in important Digitization of Cultural Heritage projects. To emphasize the distinction between Irish culture in the English language (currently the dominant culture) and the Irish culture in the Gaelic language, I chose the latter to exhibit experience in the actual International Seminar in Bansko.

The Irish Script on Screen (ISOS) Project <<http://www.isos.dias.ie/>>, led by the Dublin Institute for Advanced Studies (DIAS) illustrates very quickly the two cultures side by side. The contribution of Trinity College Dublin is exhibited at <<http://www.isos.dias.ie/english/index.html>>. Most noteworthy about the project is the bringing together digitally of manuscripts physically distributed.

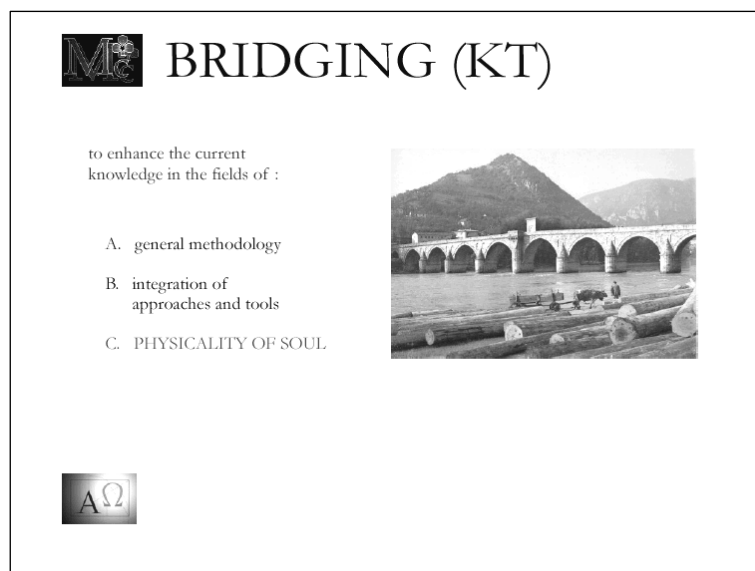
Trinity College Dublin is perhaps better known for its facilitation of the Digitization of the **Book of Kells** (written c. 800 CE). Curiously, it is a sign of the times that commercialization is given pride of place <<http://www.tcd.ie/Library/>> and <<http://www.bookofkells.ie/>>. It is regrettable that there is no online access to the Book of Kells, an unnecessary imprisonment to one place on a small and insignificant island in the globalized digital world. This gives particular meaning to the notion of being **insular** in 2004.

Finally, it seemed important to note that there was also substantial digitization of cultural heritage going on elsewhere in Ireland. Most notable is Corpus of Electronic Texts (CELT) at UCC <<http://www.ucc.ie/celt/>>.

---

## Bridging

---



Slide No.7 [of 11]

Let us **imagine** a picture [slide 7] of the Bridge on the Drina (at Visegrad) [Bridge\_Drina, 2004]. Intentionally, the bridge was not named (БЕЗ ИМЕ). But it would be noticed by the audience that pictures of bridges seemed to be an important feature of the play(ing). An interesting literary allusion (unintentional ?) to a **language-game** [Wittgenstein, 2001] appears in the book *The Bridge on the Drina* [Andric, 2000, 63] where


“Mastro Antonio's assistant, the Arab, rushed impatiently to the spot [where a great rectangular stone is being lowered onto the bridge under construction] and began with loud angry cries (in that strange composite language which had been evolved in the course of years between these men from all parts of the world) to give orders to those handling the crane on the waters [of the Drina] below.”

For me, my focus on KT–DigiCult–BG was, as I have emphasized, on multi-culturalism. It is natural to suppose that in the [ **DrDC** ] one would re–discover, re–create, one's one culture. But, in Ireland, we have two cultures side by side, one influencing the other (Yeats, Beckett, Heaney, ...). For me, it seems natural to examine how [ **DrDC** ] might be about the re–discovering, re–creating, the **culture of the other**. To attempt such a thing is what I mean by **bridging**.

In practice, two cultures side-by-side, have often been and may still be "at war". Not only may language separate or divide but different religions may go hand in hand with each culture, furthering the sense of separation. In Ireland, it is *historically* the case that the English and Irish cultures were at war and were also distinguished by differing versions of Christianity: Protestantism and Catholicism, respectively. At a local level, physical bridges have often been symbols of division. The images of bridges featured: Mostar and Drina share something of this physical bridging characteristic.

The goal of the current research is to find a way of bridging such cultures through [ **DrDC** ]. One natural way of achieving this seems to me to be by **playing together**.

## Playing




# PLAYING

A. HOMO LUDENS:  
a study of the play  
element of culture (Huizinga)  
... хоро

B. LANGUAGE-GAME  
(Wittgenstein)  
... червен (CHERVÉN)

C. EDUTAINMENT  
(Sotirova)  
... Аџа

D. IMAGE-GAME  
(Mac an Airchinnigh)



WHITE TO MOVE



Slide No.8 [of 11]

Let us **imagine** a picture [slide 8] of a chess board showing a legal configuration with "white to move". How does one know that the configuration is legal? The image itself was screen-grabbed from an actual game played between the author (white) and the eMac computer (black). Notice again the mention of the colour **white**.

Games such as chess have well-defined rules and a well-defined goal. What sort of [ **DrDC** ] game might one want to play? What are its characteristics?

As a direct consequence of coming into contact with research on EduTainMent [Sotirova, 2004a] I discovered that there was a seminal work on **playing**: "HOMO LUDENS: a study of the play element of culture" [Huizinga, 1955]. A second paper on the subject of EduTainMent [Sotirova, 2004b] reminded me of the importance of the concept of **language-game** [Wittgenstein, 2001]. I consequently wondered whether one might introduce the idea of **image-game** [Mac an Airchinnigh, 2004b]. Some preliminary experimental work has already been conducted. The Bansko performance reported here is one such very **crude** experiment and this paper (script) itself is an ongoing part of that play(ing). The images have already been presented in public, are available on the WWW, and are now interpreted for the interested players. An attempt was made to invent appropriate working rules for a [ **DrDC** ] game in a multicultural setting. In the discussion context below, the cultures are Bulgarian and Irish Gaelic with a **common communication language**: English.

## Ideas for Rules of a 2-player [DrDC] Game

*Play may be defined as an activity which is essentially free, separate, uncertain, unproductive, governed by rules, and make-believe.* Paraphrased from [Caillot 1961, 9].

Following [Salen & Zimmerman, 2004, 139] I will classify rules into 1) operational, 2) constitutive, and 3) implicit rules. Operational rules are those which explain how the game is played in practice. Constitutive rules are the



abstract mathematical and logical rules which underlie every formalizable/computable game. Finally, the implicit rules are those which cover everything else, such as good behaviour, those things which are understood in a culture, etc. It is also convenient to use the convention of naming two players of a game (**(A)**lice and **(B)**ob). Such a pair of players is conveniently referred to as a couple. Now let us **imagine** that A is Bulgarian, B is Irish and B moves first. The goal of the game is to **surprise** [ not shock! ] the other player by something in their own culture by **pointing to** a collection of WWW pages which tell a story. It is difficult to say exactly (i.e., define) what a surprise might mean. But, it entails the notion of the unexpected which gives pleasure. The nature of the [ **DrDC** ] game is special in that the pleasure of the surprise (i.e., a winning move) is a mutual pleasure for the couple and therefore the game being non-antagonistic is mutually culture bridging/building/supporting. It is easy to see how the game might be ruled/played to have the opposite devastating effect. This is one major area of concern and which needs to be addressed in **implicit** rules which are written down and agreed to in advance. Let me try to give some outline of the operational rules that have been considered so far.

One makes a move by searching the WWW (using Google, for example) and selecting up to 3 **challenges** based on WWW pages that one thinks might cause the other player to declare surprise. The 3 challenges ought to present some image or picture in A's culture that tells a story (i.e., exhibits a common theme, such as colour) that causes A to be surprised. Every move in the game is a playful step to get to know the other as game-time passes. Here is a brief outline of the state of the rules at the time of the performance in Bansko 2004. Explanatory comments are enclosed in brackets.

**R1.** The starting point of a move is the use of a search word which is considered to be a **root** word for the theme. [ For B's move he might start with прозорец (window). This is a good general word that suggests many possibilities, including for example the window of a computer and hence also a window into the WWW itself. *Caution:* single words may give too many hits. At the time of writing прозорец gives 134,000 hits. The initial game was played in the context of text search. Image search is also recommended. ]

**R2.** One can use any natural language equivalent of the root word. (fuinneog, fenêtre, ... ). [ The idea was to allow for the possibility of using culture words outside the couple's basic pair of languages. Implicitly, it meant that each player needed to have a good online dictionary for such multi-cultural play — online Bulgarian-Irish dictionary ? None exists ? ]

**R3.** Searching can only be done by Google (or any other fixed search engine).

**R4.** One can add words to a root word (but no more than 2, 3 words suffice for meaning frame). [ The primary reason for adding words is to cut down on the number of hits with a single root word. For example, прозорец цвят gives 2,450 hits; "прозорец цвят" gives 0 hits; "цветен прозорец" gives 759 hits.]

**R5.** A new root word can be chosen by the person admitting изненада (surprise). [ This signifies the ending of a winning move by B in the game and change of role play to A. ]

**R6.** The new root word is the root of the изненада. [ This signifies that the theme chosen by the first player B is worth playing on. Failure to play on is the end of that particular game. ]

**R7.** At most 3 challenges can be presented in any one move. [ Using прозорец the challenges might be  
1. <http://svishtov.com/izp/>, and with прозорец бял

2. <http://chm.moew.government.bg/pa/final.php?viewentry=704> ]

**R8.** At most 2 WWW sites can be presented in any one challenge.

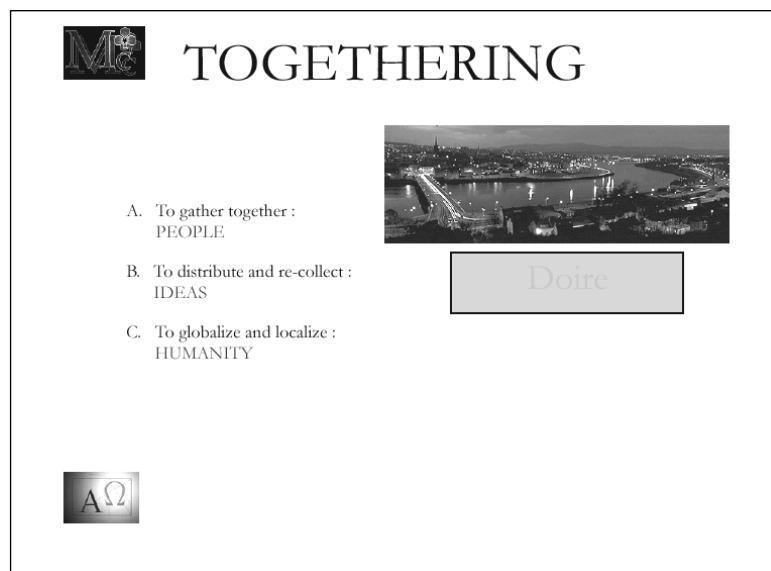
Since the carrying out of this experiment and based on subsequent research since Bansko 2004, it seems appropriate to note the following results.

**Those things which are obvious, blind one.** Each person growing up within their own culture (language, music, dance, ...) is like a tree in the forest. Everything seems as it ought to. It is only when an outsider comes along and "says" that there are trees just like you but in different kinds of forests on the other side of the world that you are shaken. You do not believe such a thing unless the outsider tells you something about yourself (and your culture) that you never noticed before, something that surprises you.

Rules for a [ **DrDC** ] game are very hard to design and formulate. That is no surprise. The above rules are very crude. Upon advice, I had seriously considered not including them at all. But I then realized that there must be some written record in public to show that an attempt had already been made. The real difficulty is to know how to

embody the playing experience of the game. By this I mean to know how the game is to be realized by **physicality of soul**. A game concerning colours ought to lead to practical colourful outcomes that can be shared – a painting perhaps ? A game concerning dance ought to lead to new or re–newed dancing experience by extension of known dancing tradition, such as a modernization of xopo ? The phenomenon of Riverdance exemplifies the general idea in the Irish Gaelic tradition. Not many know that this took place among the “Irish in the USA” and not in Ireland.

## Togethering



Slide No.10 [of 11]

Let us **imagine** a picture [slide 10] of a bridge. This bridge is in the city of Doire (in the place called Northern Ireland). The English name for the city is Londonderry and abbreviated to Derry by the people of the Irish Gaelic tradition. Doire and Derry both “mean” Oak tree. I write this text in a county called Kildare, or Cill Dara in the Gaelic, and Dara “means” Oak tree. The (hi)story of the oak trees of Ireland is a sign of deep separation and division between two cultures. How might one bridge such deep division?

**To gather together people** : the goal of the [ DrDC ] game is to bridge the deepest of divisions (linguistically or religiously or politically or ...) between two cultures. The KT–DigiCult–BG project is a “knowledge transfer” project into Bulgaria. But such a transfer will be “colonizing” and “patronizing” unless there is a reciprocal “knowledge transfer” out of Bulgaria. For me and many others in this project, this reciprocity will naturally be an embracing of (at least an understanding of) Bulgarian culture, digital and embodied in practicality of soul experience. The togethering of people of diverse cultures in a linguistic setting is likely to rely on the hegemonic Americano–Anglican “English” language for at least another generation. This text is an indicative illustration. It is therefore imperative that considerable effort be put into the exhibition of the “non–dominant” cultures participating in the togethering. The three poems БЕЗ ИМЕ (without name) were intended to emphasize this in the play(ing).

**To distribute and re–collect ideas** : the WWW permits the distribution of ideas and texts and images to all parts of the world. There is the reciprocal inverse operation of bringing back together ideas and text and images to one place, to the individual in her/his cultural setting. One expert working on a text of 100 pages will take 100 units (day, weeks, ...) in which to finish the task alone. 100 experts distributed throughout the world working on the same text will take 1 unit. The togethering might add another 10 or 20 units. Such work–togethering is the promise for our cultural heritage future in the field of digitization. The paper by Kiril Ribarov in this volume is a practical illustration of such togethering.

**To globalize and localize humanity** : But what exactly is the point of all this digital preservation of cultural heritage ? What is the purpose of the digitization of the Book of Kells, of medieval Slavonic Manuscripts, of ... ?

This is the same problematic question as "What is the purpose of Museums and Libraries?" Every people and state has a good sound reason for preservation and passing on of their cultural tradition/heritage and this is always a matter of passing on to the young, to the next generation. This is always, inevitably a matter of teaching, of awakening. It is with this idea that I finished the (play)ing performance and so here is the final scene.

### Epilogue: Why me? Будител



## WHY ME ? БУДИТЕЛ ?



"This collection ends with an essayistic question mark: the paper of **Dr. Mícheál Mac an Airchinnigh** from Trinity College, Dublin. It puts many open questions. We have not choose to end with such a provoking text because of the idea that digitisation is still *terra incognita* where one does not have a clue what has to be done and how to do it. In this field, it is of utmost importance not only to present practical work done on a certain level of quality, but to find new ways to re-present and re-arrange our past through the new technologies. The paper of Dr. Mac an Airchinnigh raises these important philosophical issues in an untraditional manner presenting scientific argument in a poetic form. This appeals strongly to us, because in our work all of us have to find the right combination of practical, earthly approaches – this is our *terra*, which is yet *incognita* because we still explore the best ways to deliver content, and the content itself, and still have no answers to all questions *How? Why? In what form? When? Who? For Whom? Where?*" [Dobrevá & Ikonomov, 2004]

(Dobrevá, Milena & Ikonomov, Nikola. *Preface. Review of the National Center for Digitization*, IV-3, 2004. ISSN 1820-0109)



## WHY US ? — БУДИТЕЛИ ?

Slide No.11 [of 11]

Let us **imagine** a picture [slide 11] of the author again. All players, actors, performers, look to others to speak of them. This is often done in a review. Sometimes the reviews are good, sometimes not. Here is a review:

"This collection ends with an essayistic question mark: the paper of **Dr. Mícheál Mac an Airchinnigh** from Trinity College, Dublin. It puts many open questions. We have not choose to end with such a provoking text because of the idea that digitisation is still *terra incognita* where one does not have a clue what has to be done and how to do it. In this field, it is of utmost importance not only to present practical work done on a certain level of quality, but to find new ways to re-present and re-arrange our past through the new technologies. The paper of Dr. Mac an Airchinnigh raises these important philosophical issues in an untraditional manner presenting scientific argument in a poetic form. This appeals strongly to us, because in our work all of us have to find the right combination of practical, earthly approaches – this is our *terra*, which is yet *incognita* because we still explore the best ways to deliver content, and the content itself, and still have no answers to all questions *How? Why? In what form? When? Who? For Whom? Where?*" [Dobrevá & Ikonomov, 2004]

To make a serious technological and scientific case through the medium of the arts and humanities is a difficult challenge. Every teacher and **будител** (awakener) knows this as performer. (S)he is like the bird which feeds its young. The adult bird eats and then disgorges the food for the young birds that hungrily wait and demand. Every teacher, actor, performer knows why this is really the only thing that works. In the case of humans, not everything is accepted or understood. Even the [ **DrDC** ] game proposed which seems to hold out so much promise, can be used to obtain exactly the opposite of mutual cultural understanding.

But the "**Why me?**" is **not** about me at all. It is a classical rhetorical question which is intended to challenge each one (in the audience then and now in the readership) who supposes that the work they do in the Digitization of Cultural Heritage has real meaning for the people of their own culture and even more importantly for the people of an "associated" culture. Ultimately, each is challenged to look to their young, to the future generations. I hope the questioning, the performance, the play(ing) will lead to new approaches, one of which I recommend and continue to explore: EduTainMent, and in particular the potential of the multi-cultural [ **DrDC** ] game.

## Acknowledgements

I would like to thank Vassil Nikolov, Aibhín O'Connor and especially Kalina Sotirova for reading an earlier draft and making substantial suggestions for improvement. A special debt of acknowledgement is due to Milena Dobрева who has kept faith with my strange approaches and ideas since 1991. The University of Dublin, Trinity College, still adheres to an ancient academic tradition which allows a mathematician and computer scientist to work in diverse fields and to cross many boundaries. To all those who have been responsible for allowing me **to do** this work and **to teach** to others, to students at home and to "students" abroad, there are no words that suffice to express my gratitude ... other than these words I write.

## Bibliography

- [Andric, 2000] I. Andric. *The Bridge on the Drina*. Dereta Publishers, Belgrade, 2000. L. F. Edwards (trans.) from Serbo-Croat. <<http://www.dereta.co.yu>>. Date of last access: 2004-10-06.
- [Bridge\_Drina, 2004] <<http://www.prevodi-vertalingen.com/bridgesformostovi/othersincandb/cuprija5.html>>. Date of last access : 2004-10-06.
- [Bridge\_Mostar, 2004] <<http://www.reuters.co.uk/newsPackageArticle.jhtml?type=topNews&storyID=552722&section=news>>. Date of last access : 2004-08-22. No longer accessible : 2004-10-06. Alternative (images) at <<http://www.prevodi-vertalingen.com/bridgesformostovi/mostar/mostar1.html>>. Date of last access : 2004-10-06.
- [Caillot, 1961] R. Caillot. *Man, Play and Games*. University of Illinois Press, Chicago, 2001. Meyer Barash (trans.) from the French *Les Jeux et les hommes*, Librairie Gallimard, Paris, 1958. ISBN 0-252-07033-X.
- [Dobрева, 1994] *Applications of Computer Tools in Studying Medieval Slavonic Manuscripts*. Institute of Mathematics, Bulgarian Academy of Sciences, Sofia, September, 1994.
- [Dobрева & Ikonov, 2004] M. Dobрева & N. Ikonov. Preface. Review of the National Center for Digitization, IV\_3/2004, 2004. ISSN 1820-0109. <[http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index\\_e](http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index_e)>. Date of last access : 2004-09-30.
- [Huizinga, 1955] J. Huizinga. *HOMO LUDENS, a study of the play element in culture*. The Beacon Press, Boston, 1955. The original work was published in 1938. Huizinga himself insisted that the subtitle in English be 'The Play Element of Culture', Foreword, Leyden 1938. The translator disagreed. ISBN 0-8070-4681-7.
- [Kane 2004] P. Kane. *The Play Ethic*. Macmillan, London, 2004. ISBN 0-333-90736-1.
- [Mac an Airchinnigh, 1991] *On the Representation and Reconstruction of Chester Beatty Papyri*, Preliminary Report (unpublished manuscript). Department of Computer Science, Trinity College, Dublin, May 1991.
- [Mac an Airchinnigh, 2004a] M. Mac an Airchinnigh. The practical sense of philosophizing : why preserve anything at all, even digitally ? Review of the National Center for Digitization, IV\_3/2004, 2004. ISSN 1820-0109. <[http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index\\_e](http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index_e)>. Date of last access : 2004-09-30.
- [Mac an Airchinnigh, 2004b] M. Mac an Airchinnigh. The graven image : digitized and philosophized ? To appear in Review of the National Center for Digitization. ISSN 1820-0109. <[http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index\\_e](http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index_e)>. Date of last access : 2004-09-30.
- [Salen & Zimmerman, 2004] K. Salen & E. Zimmermann. *Rules of Play, Game Design Fundamentals*. The MIT Press, London, England, 2004. ISBN 0-262-24045-9.
- [Sotirova, 2004a] K. Sotirova. *Edutainment Games – Homo Culturalis vs Homo Ludens*. Review of the National Center for Digitization, pp.84-98. ISSN 1820-0109. <[http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index\\_e](http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index_e)>. Date of last access : 2004-09-30.
- [Sotirova, 2004b] K. Sotirova. *Edutainment (Game) – Digital (re)Discovery of Culture*. To appear in Review of the National Center for Digitization. ISSN 1820-0109. <[http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index\\_e](http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/index_e)>. Date of last access : 2004-09-30.
- [Sutton-Smith, 1997] B. Sutton-Smith. *The Ambiguity of Play*. Harvard University Press, London, England, 1997. ISBN 0-674-00581-3. The 2001 paperback edition is cited.
- [Wittgenstein, 2001] L. Wittgenstein. *Philosophical Investigations*. Third Edition, Blackwell Publishing, Oxford, 2001. First edition 1953, Second edition 1958. ISBN 0-631-23159-5.

## Author Information

**Mícheál Mac an Airchinnigh** – Department of Computer Science, University of Dublin, Trinity College, Dublin 2, Ireland; e-mail: [mmaa@cs.tcd.ie](mailto:mmaa@cs.tcd.ie)

---

## THE EXPERIENCE OF THE ARNAMAGNÆAN INSTITUTE, COPENHAGEN

Matthew Driscoll

*Abstract: The Arnamagnæan Institute, principally in the form of the present writer, has been involved in a number of projects to do with the digitisation, electronic description and text-encoding of medieval manuscripts. Several of these projects were dealt with in a previous article 'The view from the North: Some Scandinavian digitisation projects', NCD review, 4 (2004), pp. 22-30. This paper looks in some depth at two others, MASTER and CHLT.*

*The Arnamagnæan Institute is a teaching and research institute within the Faculty of Humanities at the University of Copenhagen. It is named after the Icelandic scholar and antiquarian Árni Magnússon (1663-1730), secretary of the Royal Danish Archives and Professor of Danish Antiquities at the University of Copenhagen, who in the course of his lifetime built up what is arguably the single most important collection of early Scandinavian manuscripts in the world, some 2,500 manuscript items, the earliest dating from the 12th century. The majority of these are from Iceland, but the collection also contains important Norwegian, Danish and Swedish manuscripts, along with approximately 100 manuscripts of continental provenance. In addition to the manuscripts proper, there are collections of original charters and apographa: 776 Norwegian (including Faroese, Shetlandic and Orcadian) charters and 2895 copies, 1571 Danish charters and 1372 copies, and 1345 Icelandic charters and 5942 copies. When he died in 1730, Árni Magnússon bequeathed his collection to the University of Copenhagen. The original collection has subsequently been augmented through individual purchases and gifts and the acquisition of a number of smaller collections, bringing the total to nearly 3000 manuscript items, which, with the charters and apographa, comprise over half a million pages.*

---

### Projects

---

Following its constitutional separation from Denmark in 1944, Iceland petitioned for the return of Icelandic manuscripts in Danish repositories. After much debate, it was finally agreed that a significant part of the Arnamagnæan Collection (1666 items, in addition to all the Icelandic charters and apographa), should be transferred to Iceland, along with a smaller number of Icelandic manuscripts (141) from the Royal Library in Copenhagen, to be housed in a sister institute set up expressly for that purpose. The first two manuscripts were handed over in 1971, immediately after ratification of the treaty, and the last two in June 1998. At about that same time the Arnamagnæan Institute (in close cooperation with its sister institute in Iceland) began working towards reuniting the collection virtually through digital technology. Outlined below are some of the projects and initiatives in which the institute has become involved as a result.

#### *The MASTER project and the TEI*

The first step toward the goal of virtual reunification will be a new electronic catalogue of the entire collection, based on Kristian Kálund's *Katalog over Den Arnamagnæanske Håndskriftsamling* (Copenhagen, 1888-1894), but supplemented by more recent scholarship. Because the two Arnamagnæan Institutes are primarily research institutes, whose chief function is to further the study of the manuscripts in the collection, our records tend to be fuller than most ordinary catalogue entries, and contain occasionally quite detailed descriptions of palaeography and orthography, illumination and bindings, as well as full transcriptions of marginalia and accompanying material, such as the notes made by Árni Magnússon and his amanuenses, which are generally kept with the manuscripts to which they refer.

Preliminary work on this catalogue was undertaken as part of MASTER (Manuscript Access through Standards for Electronic Records), an international project with funding from the Telematics for Libraries section of the European Union Fourth Framework programme whose goal was to define and implement a general purpose standard for the description of manuscript materials using TEI-conformant SGML/XML (the project website, unfortunately now rather out of date, is: <http://www.cta.dmu.ac.uk/projects/master/>). The project period ran from January 1999 through June 2001. Full partners, in addition to the Arnamagnæan Institute, were: Centre for

Technology and the Arts at De Montfort University, Leicester (UK), Koninklijke Bibliotheek, Den Haag (NL), L'Institut de recherche et d'histoire des textes, Paris/Orleans (FR), The Humanities Computing Unit, Oxford (UK) and Národní knihovna České republiky, Praha (CZ). Associate partners included several major European libraries, notably The British Library (UK) and Biblioteca Apostolica Vaticana (VA), as well as a number of smaller institutions such as our sister institute in Iceland, Universitetsbiblioteket, Lund (SE), Народна Библиотека "Св Св Кирил и Методий" and Институт по Математика и Информатика, БАН, София (BG), and Lietuvos nacionaline Martyno Mazvydo biblioteka, Vilnius (LT). An independent expert group, made up of Dr Ian Doyle, Durham (UK), Professor Peter Gumbert, Leiden (NL) and Dr Gilbert Ouy, Paris (FR), monitored and commented on the development of the standard from the start. Since the end of the project period there has also been significant input from users of MASTER, which number in the hundreds, if not thousands.

MASTER had close contacts with several other projects with similar or complimentary goals: in North America the EAMMS project (Electronic Access to Medieval Manuscripts), a collaboration between the Hill Monastic Manuscript Library at Saint John's University in Minnesota and the Vatican Film Library at Saint Louis University, funded by the Andrew W. Mellon Foundation, and Digital Scriptorium, a collaboration between the Bancroft Library at the University of California at Berkeley and Columbia University's Rare Book and Manuscript Library, also funded by the Mellon Foundation; and in Europe MALVINE, funded under the same EU call as MASTER, but focusing on modern literary manuscripts and letters. Finally, the development of the MASTER document type definition (DTD) for manuscript description proceeded in tandem with the Text Encoding Initiative's Medieval Manuscripts Description Work Group (1998-2000), chaired by Consuelo Dutschke, Curator of Medieval and Renaissance Manuscripts at the Rare Book and Manuscript Library at Columbia University and Ambrogio Piazzoni, prefect of the Vatican Library.

More or less as a direct result of this, the institute became a member of the Text Encoding Initiative itself. The TEI is an international and interdisciplinary standards project established in 1987 to develop, maintain and promulgate hardware- and software-independent methods for encoding humanities data in electronic form. The TEI began as a research effort cooperatively organised by three scholarly societies (the Association for Computers and the Humanities, the Association for Computational Linguistics and the Association for Literary and Linguistic Computing), and funded by substantial research grants from, among others, the US National Endowment for the Humanities, the European Union, the Canadian Social Science Research Council and the Mellon Foundation. In December 2000 an independent and self-sustained non-profit consortium was set up to maintain and develop the TEI standard. There are currently 81 members of the TEI Consortium, including universities, research libraries, academic and other non-profit publishers, scholarly societies and others concerned with the design, production or delivery of structured electronic text (see the TEI's website: <http://www.tei-c.org>). The technical work of the TEI is overseen by an elected council, on which I have served since 2000.

The next version of the TEI Guidelines, "P5", scheduled for release in the early part of 2005, will contain a major new chapter on manuscript description, based chiefly MASTER and the work of the TEI workgroup, but with significant input also from the Repertorium of Old Bulgarian Literature and Letters project, based in Sofia and Pittsburg. I chaired the TEI task-force whose job it was to reconcile these various schemes and produce a single system for incorporation into the TEI. There remains a number of areas in need of further work, however, which will be dealt with by a properly constituted work-group, of which I shall probably also be the chair.

Meanwhile, work on the Arnamagnæan catalogue continues. During the MASTER project period itself some 500 records, the majority of them minimal, were produced in Copenhagen. It was decided to concentrate on the medieval manuscripts in the collection, although post-medieval manuscripts of special importance (for example copies of medieval vellums now lost) were also described. Since the end of the period minimal records – comprising little more than shelfmark, date and place of origin and an identification of the contents – were made for the remainder of the collection, but little more than that has been done owing to lack of funds. In Iceland basic cataloguing began in the year 2000. It was decided to include all information regarding each manuscript from the printed catalogue, translated from the original Danish to Icelandic, but, in the initial stages, no more than that. Two full-time employees carried out most of this work. All the manuscripts in the Icelandic part of the collection have now been catalogued in this manner and work has begun on "complete cataloguing", where each manuscript is examined and its contents and appearance described in detail.

The cataloguers at both institutes have produced manuals outlining the methods and terminology for cataloguing. Furthermore, the cataloguers in Iceland have in cooperation with the Icelandic software company Raqoon

designed a markup language for manuscript images (MIML) and used semantic web technology on some of the MASTER records made there (see <http://www.raqoon.com/>).

Although much has been done, much still remains to be done. The manuscripts catalogued thus far have been predominantly West-Norse (Icelandic and Norwegian), and chiefly literary; more experience is needed with Danish and Swedish manuscripts and manuscripts in Latin, and on other types of primary sources, such as diplomas.

#### *CHLT*

Another project in which the institute has become involved is CHLT (Cultural Heritage Language Technologies), a collaborative project involving other institutions in Europe and the United States: Department of English, University of Missouri at Kansas City (USA), The Perseus Project, Tufts University (USA), Department of Scandinavian Studies, University of California at Los Angeles (USA), The Newton Project, Imperial College, London (UK), Classics Department, Cambridge University (UK), Istituto di Linguistica Computazionale, Pisa (IT) and the Max Planck Institut, Berlin (DE). Funding for the project is provided by the National Science Foundation in America and the European Union International Digital Library Collaborative Research Programme. The project period runs from 1 June 2002 to 31 May 2005. The project has three major goals: first, to adapt discoveries from the field of computational linguistics and information retrieval and visualization in ways that are specifically designed to help students and scholars in the humanities advance their work; second, to establish an international framework with open standards for the long-term preservation of data, the sharing of metadata, and interoperability between affiliated digital libraries; and finally, to lower the barriers to reading Greek, Latin and Old Norse texts in their original languages (for more information see the project website: <http://www.chlt.org>).

The principal role of the Arnarnagnæan Institute in the project is the provision of electronic texts, while our American partner, the Scandinavian Department at UCLA, handles the processing of these texts, in particular the development of a morphological analyser for Old Norse. This work has been carried out by a team of very capable students, under the direction of myself in Copenhagen and Prof. Timothy Tangherlini in California. It was decided to use eight of the Fornaldarsögur Norðlanda or mythical-heroic sagas, which deal with the early history of Scandinavia, as a test corpus, basing our texts each on a single manuscript, normally the oldest but in any case the one deemed to be the best. All the texts are marked up using TEI-conformant XML. The transcriptions have generally been made on the basis of a printed edition, but as few of the extant editions reproduce the text of the original manuscripts as diplomatically as we wanted, a good deal of "un-normalisation" has been necessary. At the same time, a fully normalised form of every word is added to the mark-up for search and processing purposes.

In brief, the transcription conventions we have employed are as follows: The text is transcribed exactly as it is in the manuscript with respect to orthography and spacing between words. Variant forms of the same letter (allographs) are not distinguished, apart from small capitals, used to denote geminates (principally N and R, but potentially also D, G, M, S and T), high and round s, ordinary and round r (r-rotunda), ordinary and insular forms of f and v, ordinary and uncial forms of d, e, m and t, all of which are represented using entity references. Only ligatures with an independent phonemic value (a and e, double a etc.) are represented; ligatures which are the result of graphic economy are treated as two separate characters (high s + t, for example). Abbreviations are expanded in accordance with the normal spelling of the scribe in question, using <expand> to indicate supplied letters, and the means by which the abbreviation is achieved, i.e. the sign or tittle used expressed as an entity reference, is given as the value of the abbr attribute. Abbreviation by suspension is distinguished from abbreviation by other means (contraction, supraliner symbol etc.) by means of the type attribute so that these may be processed differently. Letters or words assumed to have been inadvertently omitted by the scribe (which in a printed edition would normally be placed in angle brackets) are supplied and tagged using <supplied reason="omitted">, while <supplied reason="illegible"> is used to indicate letters now unreadable but assumed originally to have been in the manuscript (which in a printed edition are normally placed in square brackets). Where necessary to the sense, emendations and alterations are made to the text; obvious misspellings, for example, are corrected using <corr>, with the original reading given as the value of the sic attribute. Additions and deletions made in the manuscript by the scribe or in another hand are indicated with the <add> and <del> elements. Line-, column- and page-boundaries are indicated using the empty milestone tags <lb/>, <cb/> and <pb/>, giving a number for each as the value of the n attribute. Large structural

divisions in the text, i.e. chapters, are tagged using `<div type="chapter">` and given a number. Chapter headings are tagged using `<head>`, and the nature of the `<head>`, i.e. whether it is found in the manuscript itself or supplied by an editor, indicated in the value of the type attribute. The many verses in the text are tagged using `<lg>` (line-group) for stanzas and `<l>` (line) for individual lines. Owing to the prosimetric form of much saga literature, verses normally occur within prose paragraphs; this has necessitated changing the DTD in order to allow `<lg>` to appear directly within `<p>`. Finally, each word in the text is placed inside an `<orig>` element, and the normalised form is given as the value of the `reg` attribute. Compound words written separately in the manuscript should be grouped together within a single set of `<orig>` tags, while in the opposite situation, where for example a preposition and its object are written as a single word, the two parts are treated as separate words, each placed within a set of `<orig>` tags, but with no space between them. Marks of punctuation are placed outside the `<orig>` tags. Although relatively simple, this mark-up allows for (at least) three separate views of the text – strictly diplomatic, retaining the line-breaks, variant letter forms, unexpanded abbreviations and so on of the original, semi-diplomatic or semi-normalised, where the abbreviations have been expanded and any obvious errors have been corrected, and normalised, where spelling, capitalisation, word division and so on have all been regularised – through the use of multiple style-sheets, allowing the user to decide which view he or she prefers (and the ability to toggle between them). Clicking on a word opens a window providing a translation and grammatical and morphological information, which is extremely helpful to students. We hope also to provide links to digital images of the manuscripts themselves, at least on a page-by-page, but possibly on a line-by-line basis.

---

### Author Information

---

M. J. Driscoll – Arnamagnæan Institute, Copenhagen; e-mail: [mjd@hum.ku.dk](mailto:mjd@hum.ku.dk)

## THE LATEST PRAGUE CONTRIBUTIONS TO WRITTEN CULTURAL HERITAGE PROCESSING <sup>1</sup>

Kiril Ribarov

*Abstract:* This work presents a software package ACT (Annotated Corpora of Text) for lexical and corpus processing of European written cultural sources (currently used for processing of mediaeval Slavonic manuscripts). I use ACT as a contribution towards a contextual and intelligent heritage Information Technology framework. The software is suitable for capturing characteristics of old written sources including rich language variability on word and sentential level. It is not the word-form, but its understandings/interpretations that become central processing units, which can be assigned morphology distinctions, head-words (including recensional), translation equivalents; these interpretations can be joined in multi-word units or assigned correlation to other sources. The whole annotation process is automated and individual sorting orders and morphology tags structures can easily be defined. ACT incorporates modules for: complex searches on one or more sources, creation of various ready-to-use documents, web text and image access, incorporation of lexical card-files into a corpus, and text-from-card-files reconstruction.

*Keywords:* annotation, Old-Church Slavonic, lexical processing, cultural heritage

---

<sup>1</sup> The following text has been originally published in the Proceedings of the Language Recourses and Evaluation Conference held in Lisbon, Portugal, 2004, under the title of "Towards Intelligent Written Cultural Heritage Processing - Lexical processing". I present here a revised contribution of the aforementioned paper and I add here the latest efforts done in the Center for Computational Linguistic in Prague in the field under discussion.



---

## 1. Introduction

---

I suggest that intelligent heritage IT framework should place the written cultural sources in an electronic contextual (e-context) field with two major connecting elements:

- (a) source image along with language based contextual structure of the word mass present in the sources;
- (b) connections (inner and outer links) among various types of written cultural sources within a wider cultural environment.

Such framework incorporates technologies and tools necessary for large-scale activities aimed towards multi-aspectual presentation of written cultural heritage in a highly distributed manner.

Applied on mediæval Slavonic written cultural heritage in accordance with the above stated intelligent heritage framework, this work is aimed as an outline of:

- (1) the main functions of Annotation Corpora of Text<sup>1</sup> (ACT), a language independent<sup>2</sup> software tool for lexical and corpus processing of written cultural sources;
- (2) the language specifics implemented in ACT;
- (3) the first release of lemmatized and POS-annotated Old-Church Slavonic (OCS) language resource (LR)<sup>3</sup>.

This work is another step, hopefully forward, in series of continuous efforts in computerized language processing of Old-Church Slavonic (OCS) manuscripts, the most recent papers of which are [Camuglia, Camuglia, Ribarov, 2003], [Camuglia, Ribarov 2003], and [Ribarov, Camuglia, 2003] followed by two master thesis [Bubnik 2004] and [Celak 2004].

---

## 2. On Language Specifics

---

Apart from contemporary languages the old sources are characterized with problems relevant, among others, to the development of the language (synchronic, diachronic and diatopic characteristics), low presence of language spelling norms, as well as by influences from frequently used translations from other languages. Thus, the language problems to resolve exhibit particularities, which make the usage of current lexicographic stations or corpus managers impossible. The most important of the distinctions (particularities) are:

- scriptum continuum,
- variants at various levels of the language,
- abbreviations,
- damaged and unknown parts,
- correlation to other sources,
- multi-lemmatization (due to existence of various recension centers and high level of variability, and/or due to lack of material, usually, lemmatization means assignment of more than a single lemma),
- existence of translation equivalents important for, e.g. contents reconstruction and variability resolutions.

Along with the OCS resources the ACT system is taken as a framework capable of manipulation and capturing of the high-level language variability on word and/or sentential level.

---

<sup>1</sup> ACT is accessible via <http://ckl.ms.mff.cuni.cz/~ribarov> or <http://prometheus.ms.mff.cuni.cz/act> (further ACT web page). ACT has been developed as a student project at the Faculty of Mathematics and Physics at Charles University in Prague, Czech Republic, lead by Kiril Ribarov. The programming part has been developed by: Jiri Bubnik, Jiri Celak, Vojtech Janota, Alexandr Kara, Vaclav Novak; the web interface was developed by Tomas Vondra.

<sup>2</sup> Within the original version the language independence was restricted to linearizable, left to right languages. Latest changes allow that other languages are processed as well, e.g., Arabic. Testing with Arabic in ACT was verified in the master thesis of Jiri Bubnik [Bubnik, 2004]

<sup>3</sup> For web access to the OCS material visit <http://prometheus.ms.mff.cuni.cz/act/www>

*Some examples*

A simple example<sup>1</sup> on surface variability due to scriptum continuum would be

и||егоже||видиши||плода||се||сзтвори||вз||мнѣ  
(and the fruit you see created in me),

where the string *сзтвори||вз||мнѣ* could also be divided as *сзтвориѣз||мнѣ* (where *сзтвориѣз* is the past participle – active mood of *create*), so that both are grammatically correct, but the correct reading can be found only in a wider context. Such wider context is not always available.

Abbreviations of various types, damaged or unknown parts are very frequent and as such they introduce higher level of variability in interpretation and understanding. In order to process them, they need to be rendered, e.g.: (*сѣнѣ* → *с[ы]нѣ son*), (*глыѣте* → *гл[агол]ите say*), (*гѣ* → *г[оспод]ь God*), (*црѣ* → *ц[ѣ]с[л]рь King*), (*рѣ* → *рѣ[ч]е say*), (*придох* → *придох[з] come*).

Although for processing of the contemporary languages it is taken as granted that the main unit to process is either a word-form or a sentence (e.g. for parsing) such a priori certainty is not possible for, e.g. OCS: scriptum continuum eliminates punctuation signs<sup>2</sup> and surface sentence is impossible to capture; some uncertainties in word-form boundaries were stated above. The rendered form is understood as interpretation of the surface.

We suppose that other old language documents, as well as the OCS ones exhibit not only orthographic variability, but also morphological or syntactic one. We stress the need to design systems capable of recording variability on various levels – due to the closeness of the corpuses of dead languages any disambiguation process lacks the support of a wider language context or living language evidence in order to approve disambiguation choices.

### 3. ACT Solutions

In this part, only the most characteristic solutions will be pointed out. Those are in close relation to variability resolutions. We will present that the main processing unit is not the surface word-form, but its understandings; we will also present that the main "syntactic processing unit" is not a sentence but a set of any type of multi-word units.

#### Set of rendered word-forms

In order to resolve the word-level variability, a word-form is understood as a pair (original form, set of rendered forms). The string of characters identified as a part of an image or as a part of a text (e.g. scriptum continuum) delimited by the user or word-segmentation algorithm, represents the original-form (e.g. *сзтвориѣз*). The understanding, or the set of possible understandings of the original form is a set of rendered forms (variant 1: *сзтвори вз*, variant 2: *сзтвориѣз*). A single original form may have various rendered forms in two levels:

- horizontal: the original form is identified as series of neighboring rendered forms (as in variant 1, two rendered word-forms exist: *сзтвори вз*)
- vertical: the original form exhibits variants of the rendered forms, which are listed as alternatives such that each of them can become a part of a(n) (alternative) context.

A rendered form (further word-form, word) becomes a main processing unit, which is further:

- assigned a morphology distinction (or a set of morphology distinctions in case of an unresolved variant)
- assigned a head-word (disambiguated lemma accompanied by basic dictionary information and/or inter head-word's links) or a set of possible head-words in case of a variant; a head-word is further placed within a specific recension and linked within a network of equivalent recension head-words,
- assigned a translation equivalent (or a set of possible equivalents), if any,
- correlated to other sources, if any,

<sup>1</sup> The example is taken from the *Povest o Varlaam i Joasaf*, an unpublished manuscript stored at the Rila Monastery (Bulgaria) under the signature 3/14.

<sup>2</sup> Punctuation marks are more frequent in newer documents and may characterize tendencies of creation of, originally missing, spelling norms.

- 
- assigned a complex (or a set of complexes<sup>1</sup>, see *later*).

Recently, a new automatic word segmentation tool has been released [Celak, 2004]. This tool is able to treat certain variability (e.g., abbreviation) and can be applied on scriptum continuum rendering of Old-Church Slavonic. The tool, although developed separately, is ACT compatible.

Within user-friendly environment, assignment of morphology, of head-words and of translations links is automated in order to speed up the manual parts of annotation and lexical work as much as possible. The process of rendering, that is assignment of rendered form to an original form, is also automated through creation of ordered lists of re-writable rules based on regular expressions.

---

## Complexes

---

Any kind of multi-word unit is called a complex. The term complex is used because of the freedom to assign any kind of liberally distant link between any two (or among a set of) words. ACT supports user definable complexes, therefore complexes of various types. Each rendered form can become a member of a complex.

The possibility to determine various complex types allows the user to study the texts on various levels, and to resolve phrasal, idiomatic, and/or sentential variability. Starting from the simple ones, one may define complexes of, e.g. the following types:

- analytic verb form,
- reflexive particle,
- noun phrase,
- prepositional phrase,
- a whole sentence, if identifiable,
- discourse relation,
- idiom,
- citation,
- date, etc.

This possibility permits to treat the text as string of words with various stand-off structures above it, not restricted to spelling or other norms. The work [Bubnik, 2004] enriches the complexes for it allows their annotation, a so far unstructured tag can be assigned to any complex type.

---

## Complexes for Translations and Processing of Other Languages

---

The set of documents processed in ACT are organized in catalogues, a folder of documents with given language specifics. Various instances of a catalogue can be created, each of them, if needed, with different language specifics as character set coding, sorting order, and morphological tag structure.

Assuming that manuscripts were frequently rewritten in the past or translated from other languages (OCS are often translations from Ancient Greek or Latin) marking translation equivalents is needed for correct understanding of the, e.g. damaged part of the original document.

ACT allows establishment of translation links between documents of two different catalogues. These links are established between complexes, assuming that:

- a complex of translation type is defined,
- each word-form is a complex,
- for many-to-many translation relation the corresponding group of word-forms are marked as complexes of the required translation type.

During translation equivalents' assignment, ACT builds a translation memory, which is further used for automatic suggestion of translation pairs.

---

<sup>1</sup> Any type of multi-word unit.

## Automation and Heuristics

As mentioned earlier, ACT builds history lists. All annotation process is recorded and annotations are suggested to the user. To speed up this process probabilities are calculated over the history annotations. Thus, annotation can be done automatically (selecting the most probable candidate) or the user can be presented an ordered (by probability) list of possibilities. Further, the user may benefit from a promptly displayed word-form/lemma picture. These new probabilistic ACT features were implemented in [Bubnik, 2004].

## The DTD

During the last two years, significant developments of the original STINO, now ACT system were made in the stream of the already performed or announced changes, as in [Ribarov, 2002]. The whole original system has been reprogrammed and new data formats have been introduced<sup>1</sup>. Besides others, newly, XML format has been designed<sup>2</sup> with the below-presented DTD. This DTD is included at this point in order to state implicitly the ACT annotation span.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT bindkeyword (keyword)>
<!ELEMENT complex (#PCDATA)>
<!ATTLIST complex
    complex_group_refid IDREF #REQUIRED
    position CDATA #REQUIRED
>
<!ELEMENT complex_group (#PCDATA)>
<!ATTLIST complex_group
    complex_type_refid CDATA #REQUIRED
    refid IDREF #REQUIRED
    note CDATA #IMPLIED
>
<!ELEMENT complex_groups (complex_group+)>
<!ELEMENT document (pages, originalform+, complex_groups)>
<!ATTLIST document
    created CDATA #IMPLIED
    notes CDATA #IMPLIED
    place CDATA #IMPLIED
    scannedmanuscriptdir CDATA #IMPLIED
    documentAbbreviation CDATA #REQUIRED
    date CDATA #REQUIRED
    idsorting CDATA #REQUIRED
    idredaction CDATA #REQUIRED
    dateofcreationupper CDATA #REQUIRED
    idtranslation CDATA #IMPLIED
    manuscriptfont CDATA #IMPLIED
    dateofcreationlower CDATA #REQUIRED
    name CDATA #REQUIRED
    exportType CDATA #REQUIRED
    typization CDATA #IMPLIED
>
<!ELEMENT keyword (#PCDATA)>
<!ATTLIST keyword
    partOfSpeech CDATA #IMPLIED
    id_ident IDREF #IMPLIED
    lemma CDATA #IMPLIED
    paradigm CDATA #IMPLIED
```

<sup>1</sup> All of these changes are in compliance with the basic framework principles published in my earlier works.

<sup>2</sup> For technical specification, system design or other questions see ACT documentation.

```

        homonym CDATA #IMPLIED
        refid IDREF #IMPLIED
        idredaction CDATA #REQUIRED
    >
    <!ELEMENT morphology (text)>
    <!ATTLIST morphology
        keyword_refid IDREF #IMPLIED
    >
    <!ELEMENT originalform (text, renderedform)>
    <!ATTLIST originalform
        form_image_url CDATA #IMPLIED
        row IDREF #IMPLIED
        positioninrow IDREF #IMPLIED
        page IDREF #IMPLIED
        external_id CDATA #IMPLIED
    >
    <!ELEMENT page (#PCDATA)>
    <!ATTLIST page
        user_page_part CDATA #IMPLIED
        page IDREF #IMPLIED
        page_image CDATA #IMPLIED
        user_page IDREF #IMPLIED
    >
    <!ELEMENT pages (page+)>
    <!ELEMENT renderedform (text, morphology?, complex?,
bindkeyword?)>
    <!ATTLIST renderedform
        variantnumber CDATA #IMPLIED
        colocationright CDATA #IMPLIED
        otherSource CDATA #IMPLIED
        colocationleft CDATA #IMPLIED
        renderedForm CDATA #REQUIRED
    >
    <!ELEMENT text (#PCDATA)>

```

---

## On Inputs and Outputs

---

ACT inputs can read RTF, TXT, and XML file formats. The RTF and TXT format may include characters with special meaning (mark-up characters). Any type of user defined search becomes an output written as a file or displayed on the screen. Output file formats are: HTML, RTF, TXT, XML.

The user defined searches can search for any kind of information subset relevant to a word-form (wildcard characters for any attribute values can be used), as e.g.:

- word-forms that initiate, include or end on some character,
- word-forms with some morphological features
- all word-forms of a lemma (head-word),
- word-forms of a given complex type,
- word-forms in which vicinity another word-form occurs,
- word-forms with specific translation, etc.

Any type of searches can be performed on one or more than one document, within a single catalogue. Any type of searches (including complete lists of all word-forms) can be, according to user selection, presented in a form of:

- a list
- index veborum
- retrograde index

- concordance index
- frequency list.

Any of the outputs can be sorted according to various sorting criteria. The outputs are also basic statistic-oriented outputs, as frequencies and bi-gram lists.

The searches are implemented via a query assistant, which is adaptable and can be defined by user needs.

The newest ACT input module is developed separately. The idea is to process and pre-process separately any kind of input texts and formats. [Celak, 2004] successfully accomplishes this aim. The separate input module outputs a ACT XML file, which can be safely input in ACT.

---

### **Electronic Publishing**

---

Significant piece of work on outputs and electronic publishing is presented in [Celak, 2004]. The output related modules allow creating of PDF output files based on one or more manuscripts or subparts of them based on sophisticated search query. The electronic publishing system allows that the PDF output files can be mutually inter-linked.

---

### **ACT Web**

---

The document material presented in a form of scanned collections of pictures, pages of rewritten texts, and annotated corpus can be accessed via the ACT-Web module, accessible at the address as stated in the introduction of this paper.

With its 700,000 word forms<sup>1</sup>, most of which lemmatized with assigned POS, available also in a form of a text and some of them scanned, the ACT-Web collection is a unique one and the biggest of its kind accessible in electronic form via Internet.

The ACT-Web module allows a user to:

- select a manuscript or a subset of manuscripts,
- perform a search on a part of a word-form, morphology tag, head-word,
- display results with concordances,
- display manuscript text and picture if available.

The web access is at <http://prometheus.ms.mff.cuni.cz/act/www>.

---

### **ACT for Card-Files**

---

In accordance with [Ribarov, 2002] and [Ribarov, Camuglia, 2003] ACT module, called Distiller, is, up to my knowledge, the first module for incorporation of card-files into a corpus.

By a card-file, a lexicographic card-file is understood, e.g. card-file with some subset of the following information:

- lemma (head-word),
  - additional lemma (serves for more specific definition of the lemma, usually in multi-word components),
  - word-form (obligatory),
  - morphological identification of the word-form,
  - word-form ID, location in the manuscript (obligatory)
  - correlation of the word form to other sources,
  - context of the word form (obligatory),
  - translation of the word form, including the context of the translated part.
- 

<sup>1</sup> In terms of distinct word-forms 163,607 were recorded, with 15,941 distinct lemmas. On latest and more detailed statistics on the corpus data see [Bubnik, 2004].

---

ACT Distiller permits the user to:

- view scanned card-file cards
- rewrite the obligatory parts of the cards.

Rewriting the obligatory parts of the card-files follows the following steps:

- 1 The word-form location is inserted manually (as a part of further considerations a design of OCR system for automatic location identification is planned; for notes on card-file structure see [Ribarov, Camuglia, 2003]).
- 2 Relative to the inserted notation closer and wider contexts are displayed:
  - i. if the word-form to be inserted is already in the context the user is only expected to verify the information,
  - ii. if the word-form is missing, the word-form is added together with the parts of the missing context.

The other card-file information is filled in as a part of an annotation process within the ACT main module; in this case the word-form to process (lemmatize, tag) is accompanied by the card-file image.

To ease manual check-up, ACT-Distiller incorporates a context binding tool and a comparative tool that visualizes possible overlaps, mistakes, and differences.

---

## Conclusion

Let us, therefore, conclude that: ACT integrates tools necessary for state-of-the-art linguistic processing and presentation of written cultural heritage sources, demonstrated on mediaeval Slavonic written cultural heritage sources. It contributes towards a creation of adequate and innovative intelligent heritage Information Technology framework for addressing digital presentation of written cultural sources. In general, the ACT framework does not neglect the possibilities for link establishment to other (e.g. European) written cultural sources. Along with the presented OCS LR, ACT fills in the currently existing gap in the European e-space where mediaeval Slavonic cultural heritage is presented in scattered and non-unified manner.

---

## Acknowledgements

This work was supported by the Center of Excellence, Center for Computational Linguistics, project number LN00A063 of the Czech Ministry of Education.

---

## Bibliography

- [Bubnik, 2004] J. Bubnik (2004). "Automatizované značkování (středověkých) textů-heslová slova, morfologie, komplexy, korelace", MSc Thesis, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.
- [Camuglia, Camuglia, Ribarov, 2003] G. Camuglia, M. Camuglia, K. Ribarov (2003). "Computer Processing of a Clopen Language: Old Church Slavonic", In *Linguistica Computazionale*, Volume XVI-XVII, Special Issue, Editors: A. Zampolli, N. Calzolari, L. Cignoni. Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma.
- [Camuglia, Ribarov, 2003] M. Camuglia, K. Ribarov (2003). "Old-Church Slavonic in Codes", In: *Computational Approaches to the study of Early and Modern Slavic Languages and Texts-Proceedings of the "Electronic Description and Edition of Slavic Sources"*, Pomorie, Bulgaria. Sofia.
- [Celak, 2004] J. Celak (2004). "Automatizovaná segmentace, rozepisování, a správa běžných vstupů a výstupů pro zpracování (středověkých) textů", MSc Thesis, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.
- [Ribarov, 2002] K. Ribarov (2002). "Old Sources and Modern Procedures". In: *Proceedings of LREC 2002*, Spain.
- [Ribarov, Camuglia, 2003] K. Ribarov, M. Camuglia (2003). "Incorporation of Old Church Slavonic Card Files into a Corpus", In: *Scripta & e-Scripta*, Volume 1, Institute of Literature, Bulgarian Academy of Sciences, Sofia.

---

## Author Information

**Kiril Ribarov** – Research fellow, Center for Computational Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Malostranske nam. 25, Prague 1, Czech Republic, e-mail: [ribarov@ufal.mff.cuni.cz](mailto:ribarov@ufal.mff.cuni.cz)

## DIGITISATION PROCESSING AND RECOGNITION OF OLD GREEK MANUSCRIPTS (THE *D-SCRIBE* PROJECT)

**Stavros Perantonis, Basilis Gatos, Konstantinos Ntzios, Ioannis Pratikakis,  
Ioannis Vrettaros, Athanasios Drigas, Christos Emmanouilidis,  
Anastasios Kesidis, and Dimitrios Kalomirakis**

*Abstract:* After many years of scholar study, manuscript collections continue to be an important source of novel information for scholars, concerning both the history of earlier times as well as the development of cultural documentation over the centuries. *D-SCRIBE* project aims to support and facilitate current and future efforts in manuscript digitization and processing. It strives toward the creation of a comprehensive software product, which can assist the content holders in turning an archive of manuscripts into a digital collection using automated methods. In this paper, we focus on the problem of recognizing early Christian Greek manuscripts. We propose a novel digital image binarization scheme for low quality historical documents allowing further content exploitation in an efficient way. Based on the existence of closed cavity regions in the majority of characters and character ligatures in these scripts, we propose a novel, segmentation-free, fast and efficient technique that assists the recognition procedure by tracing and recognizing the most frequently appearing characters or character ligatures.

*Keywords:* Handwriting Recognition, Character Recognition, Binarization, Segmentation-free, Feature Extraction, Historical Document Recognition, Old Manuscript Recognition.

---

### Introduction

Recognition of old Greek manuscripts is essential for quick and efficient content exploitation of the valuable old Greek historical collections. *D-SCRIBE* (<http://it.demokritos.gr/cil/dscribe/>) is a Greek GSRT-funded R&D project which aims to support and facilitate current and future efforts in manuscript digitization and processing. It strives toward the creation of a comprehensive software product, which can assist the content holders in turning an archive of manuscripts into a digital collection using automated methods. Our final product will give memory institutions the opportunity to:

- Digitize their manuscript collections according to quality metrics, leveraging existing material with state-of-the-art technical feasibility.
- Produce varying digital objects for varying purposes, e.g. access vs. preservation.
- Automate the transliteration of manuscripts, by employing manuscript-tuned OCR modules.
- Manage their content in the form of a digital library, by using a powerful document management system.
- Facilitate and expand the study of paleography, by providing self-study tools which will help students and researchers in coping with large volumes of data.

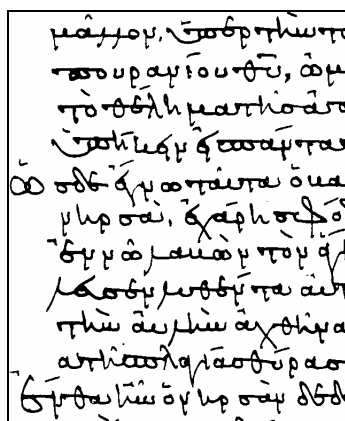
In the framework of *D-SCRIBE*, the system *STUD-IOS* will be developed. This system is constituted by two sub-systems. The first sub-system aims to cover the need for self-instruction of *D-SCRIBE* users for the digitization and treatment of Greek manuscripts. It will be developed after the completion of the *D-SCRIBE* platform, and its current planning forecasts its operation in the form of friendly support (windows help), during the operation of the *D-SCRIBE* platform. The second sub-system comes to cover the cognitive object of paleography, on theoretical issues, e.g. types of writings, faculties of paleography, materials used for the paleography, techniques of paleography. Furthermore, it supports teaching and develops the faculty of transcription of manuscripts in modern Greek.

In this paper, we focus on the problem of recognizing early Christian Greek manuscripts. Specifically, our principal concern constitutes the Sinaitic Codex Number Three, which contains the Book of Job, one of the best Greek manuscripts and one of the major masterpieces of world literature (see Fig. 1a). Written in Hebrew initially, the Book was translated into Greek approximately the 3rd century B.C. for the sake of the Hellenized Hebrews of

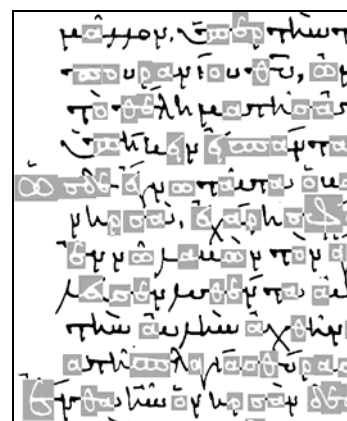


Alexandria. We propose a novel digital image binarization scheme for low quality historical documents allowing further content exploitation in an efficient way. Based on the existence of closed cavity regions in the majority of characters and character ligatures in these scripts, we propose a novel, segmentation-free, fast and efficient technique that assists the recognition procedure by tracing and recognizing the most frequently appearing characters or character ligatures.

In the field of handwritten character recognition a great progress has occurred during the past years [Vinciarelli, 2002]. Many methods were developed for a variety of applications like automatic reading of postal addresses [Lu, 2002], fax forms [Hirano, 2001] and bank checks [Xu, 2001], form processing, etc. In the literature, two main approaches can be identified: the global approach [Suen, 1993] and the segmentation approach [Kavallieratou, 2002]. The global approach entails the recognition of the whole word while the segmentation approach requires that each word has to be segmented into letters. Some approaches that do not involve any segmentation task are based on concepts and techniques that have been used in object recognition with occlusions [Chen, 2003]. According to these approaches, significant geometric features such as short line segments, enclosed regions and corners are extracted from a fully unsegmented raw document bitmap by methods like template matching [Duda, 1973], peephole method [Mori, 1992], n-tuple feature [Jung, 1996] and hit-or-miss operator [Gonzalez, 1992].



(a) Early Christian Greek manuscript



(b) Identified characters or character ligatures that contain closed cavities

Fig. 1.

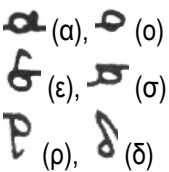
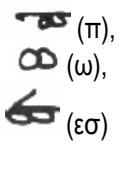
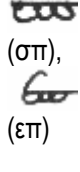
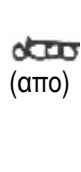
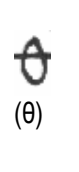
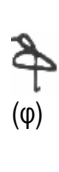
Traditional techniques for handwriting recognition cannot be applied to early Christian Greek manuscripts written in lower case letters, since continuity between characters of the same or consecutive words does not permit character or word segmentation. Furthermore, the aforementioned manuscripts entail several unique characteristics as in the following:

- High script standardization. Although, we refer to handwritten manuscripts, the corresponding characters are highly standardized since the manuscripts are immediate predecessors of early printed books.
- Frequent appearance of character ligatures
- Frequent appearance of closed cavities in the majority of character and character ligatures. As shown in Fig. 1b, closed cavities appear in letters “α”, “ο”, “σ”, “ε”, “δ”, “ω”, “π”, “θ”, “φ” as well as in letter ligatures “σπ”, “εσ” etc. These constitute 60% of complete character set used in a typical old Greek manuscript.

The continuity between characters of the same or consecutive words guided us to develop a segmentation-free recognition technique as a fundamental assistance to Old Greek handwritten Manuscript OCR. Based on the existence of closed cavities in the majority of characters and character ligatures, we propose a technique for the detection and recognition of characters that contain closed cavities. It is a novel method whose originality is based on two aspects. First, a novel segmentation-free approach based on the detection of the closed cavities. This aids toward the proposed character representation since the hole regions exist in the majority of characters and character ligatures. Second, novel features are used that are based on the protrusions in the outer contour of the connected components that contain closed cavities.

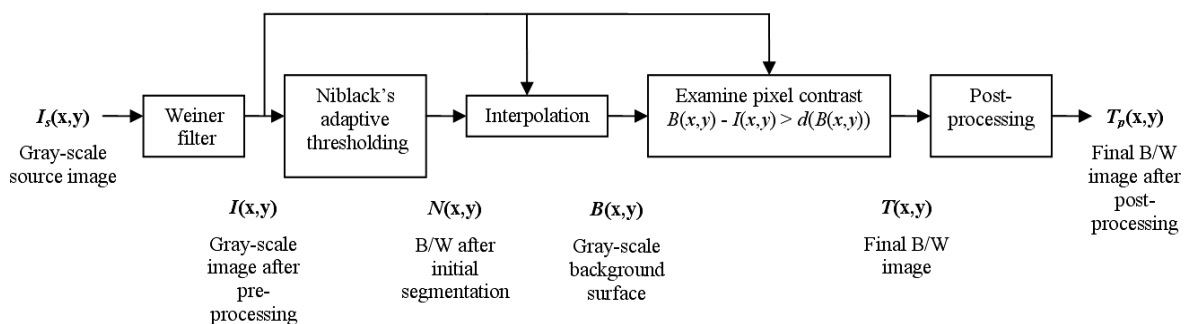
The proposed methodology consists of several distinct stages. First, we apply a binarization and image enhancement technique to get an improved quality black and white (b/w) image. Second, we trace closed cavities that exist in character bodies. We suggest a novel fast algorithm based on processing the white runs of the initial b/w image. This algorithm permits the extraction of the character closed cavities but rejects closed cavities of larger dimension, such as closed cavities inside frames, diagrams, etc. In the next step, all closed cavities in characters are initially grouped into several categories based on their spatial proximity and topology. In this way, character closed cavities are classified as: a single closed cavity, two horizontal neighboring closed cavities, three horizontal neighboring closed cavities, four horizontal neighboring closed cavities, two vertical closed cavities and two vertical neighboring patterns that consist of a single closed cavity and two neighboring closed cavities (see Table 1). The final stage of our approach concerns classification of the aforementioned closed cavity patterns into a character or a ligature. It is based on the protrusions that appear in the outer contour outline of the connected components which contain the character closed cavities. The proposed novel recognition methodology, as well as the initially applied binarization and image enhancement tasks are fully described in the following sections.

**Table 1.** The proposed dictionary for closed cavity patterns.

<b>Pattern ID</b>	1	2	3	4	5	6
<b>Pattern</b>	0	0 0	0 0 0	0 0 0 0	0 0	0 0 0
<b>Characters or character ligatures</b>	 (α), (ο) (ε), (σ) (ρ), (δ)	 (π), (ω), (εσ)	 (σπ), (επ)	 (σπο)	 (θ)	 (φ)

### Image Binarization and Enhancement

Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale image to a binary image. Since historical document collections are most of the times of very low quality, an image enhancement stage is also essential. In the literature, the binarization is usually reported to be performed either globally or locally. The global methods (global thresholding) use a single threshold value to classify image pixels into object or background classes [Otsu, 1979], whereas the local schemes (adaptive thresholding) can use multiple values selected according to the local area information [Kim, 1996]. Most of the proposed algorithms for optimum image binarization rely on statistical methods, without taking into account the special nature of document images [Niblack, 1986]. Global thresholding methods are not sufficient for document image binarization since document images usually have poor quality, shadows, nonuniform illumination, low contrast, large signal-dependent noise, smear and strains. Instead, adaptive to local information techniques for document binarization have been developed [Sauvola, 2000].



**Fig. 2.** Block diagram of the proposed methodology for low quality historical document text preservation.

The proposed scheme for image binarization and enhancement is described in [Gatos, 2004] and consists of five distinct steps (see Fig. 2): a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions using Niblack's approach [Niblack, 1986], a background surface calculation by interpolating neighboring background intensities (see Fig. 3), a thresholding by combining the calculated background surface with the original image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity. An example of the image binarization and enhancement result is demonstrated in Fig 4.

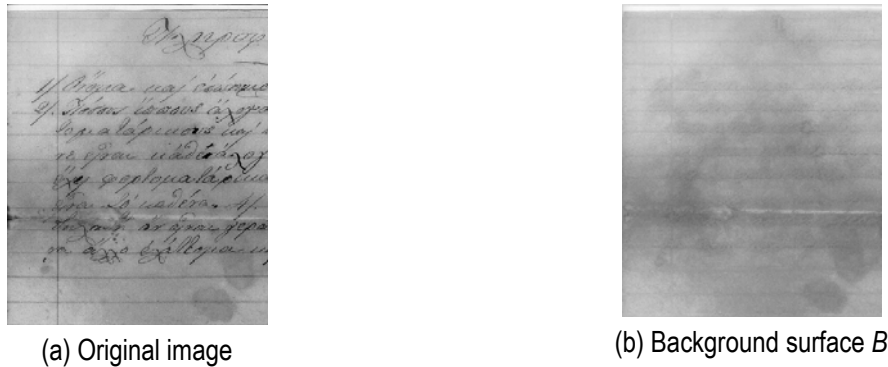


Fig. 3. Background surface estimation



Fig. 4. Image binarization and enhancement example

### Character Closed Cavity Detection

Several closed cavity detection algorithms exist, mainly based on contour following techniques that distinguish the external from internal contours [Xia, 2003]. We suggest a novel fast algorithm for closed cavity detection based on processing the white runs of the b/w image. In the following, a step-by-step description of the proposed algorithm, is given.

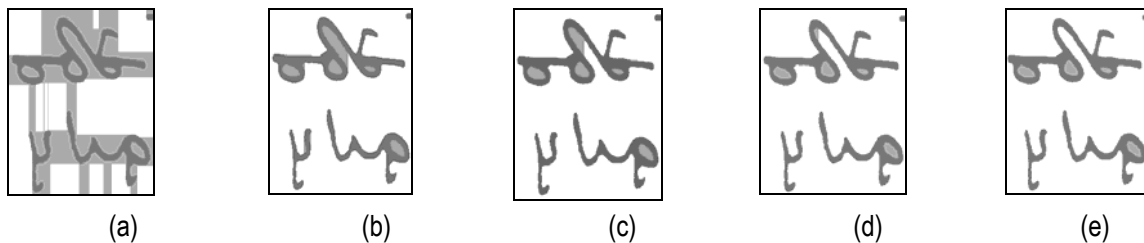
*Step 1.* All horizontal and vertical image white runs that neighbor with image borders or have a length greater than  $L$ , get flagged, where  $L$  denotes a typical length which reflects character size. The proposed algorithm for closed cavity detection extracts only the character closed cavities and not other closed cavities of larger dimension, with white run length greater than  $L$ , such as closed cavities inside frames, diagrams etc.;

*Step 2.* All horizontal and vertical white runs of non-flagged pixels that neighbor with the flagged pixels of step 1, get flagged as well;

*Step 3.* Repeat step 2 until no pixel remains to be flagged;

*Step 4.* All remaining white runs of non-flagged pixels belong to image closed cavities. Closed cavities with very small area are ignored.

An example of the proposed closed cavity detection algorithm is demonstrated in Fig.5.

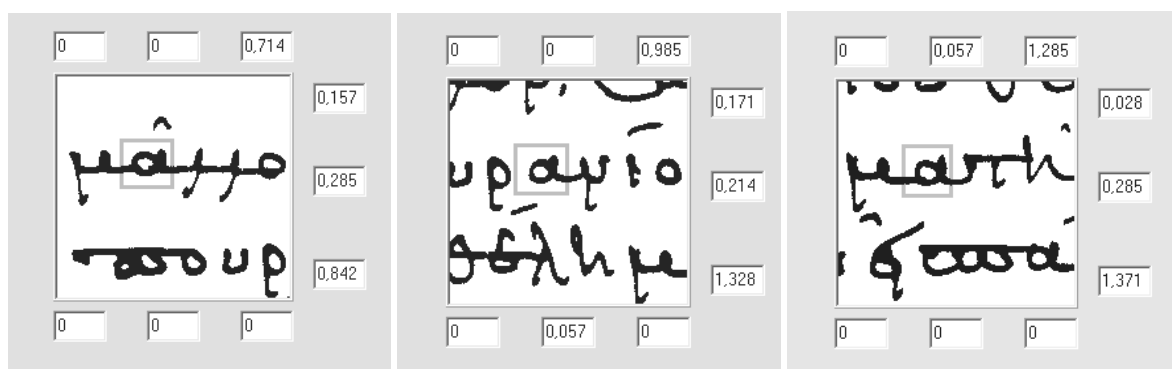


**Fig. 5.** Closed cavity detection algorithm: (a)-(e) Resulting image after 1,2,3,4 and 5 iterations, respectively.

In early Christian Greek manuscripts, a single character or character ligature may contain more than one closed cavities. Therefore, it is imperative to examine whether the detected closed cavities can be grouped together. This is done by taking into account their spatial proximity and topology leading to the construction of a dictionary. Table 1 shows the proposed dictionary structure. At the last row of this Table, we indicate the corresponding characters and character ligatures.

### Feature Estimation

Feature extraction is applied to characters that contain one or more closed cavities. The proposed method for character isolation creates a bounding box around the character guided by the spatial relationship between the contours of the closed cavity and the outer contour of the connected component. The feature estimation stage identifies all segments that belong to a protrusion in the outer contour of each isolated character. It is applied in two consecutive modes: a vertical and a horizontal mode. The vertical mode is used to describe the protrusible segments that may exist either at the top or at the bottom of the character while the horizontal mode is used to describe the protrusible segments that may exist on the right side of the character. The feature set which is composed of 9 features  $F = \{f_1, f_2, \dots, f_9\}$  expresses the probability of a segment being part of a protrusion. Features  $\{f_1, f_2, f_3\}$  describe the protrusible segments that may appear on the top of the character, features  $\{f_4, f_5, f_6\}$  describe the protrusible segments that may exist at the bottom of the character and features  $\{f_7, f_8, f_9\}$  describe the protrusible segments that may exist on the right side of the character. We have not taken into account segments that may belong to left protrusions, due to our observation that in all cases they correspond to a letter ligament rather than the main body of a character. An example of the estimated features is given in Fig. 6. One may observe a set of nine (9) cells which enable the visualization of feature values. Each cell position corresponds to the position of the respective protrusible segment in the set  $F$ .



**Fig 6.** Feature estimation example for characters "α"

### Experimental Results

The purpose of the experiments was to test the classification performance of the handwritten manuscripts with respect to the proposed closed cavity detection and feature extraction techniques. The overall experimental samples originate from three different writers of the Book of Job collection, manually labeled with the correct

answers. We have built a dictionary of closed cavity patterns that contains a total of 967 characters and character ligatures. A detailed distribution of the underlying patterns along with their spatial configuration, is shown in Table 2. We mention that the majority of characters are classified as having one or two adjacent closed cavities. Furthermore, as it is shown at Table 1, it is worth noticing that since patterns with ID 4-6 correspond to a unique character or character ligature, detection leads directly to the classification of the corresponding characters.

**Table 2.** The dictionary of closed cavity patterns including the number of pattern occurrences.

<b>Pattern ID</b>	1	2	3	4	5	6
<b>Pattern</b>	o	oo	ooo	oooo	o	o
<b>Occurrences</b>	786	130	7	3	30	11

The first set of experiments tested the performance of the closed cavity pattern detection algorithm. We recall that the dataset involved in our experiments have been preprocessed with a binarization and image enhancement algorithm. Due to this, we have worked on images of improved quality (see Figure 4).

The closed cavity detection algorithm requires no training and all of the available labeled samples were used as a test set. Table 3 shows the results obtained by applying the algorithm, indicating the recall and the precision rates for each one of the closed cavity patterns. Recall is the number of correct closed cavities found divided by the total number of existing closed cavities. Precision is the number of correct closed cavities found divided by the total number of closed cavities found. As seen from Table 3, the performance on both recall and precision is satisfactory.

**Table 3.** Recall / Precision for the characters or character ligatures in each of the closed cavity patterns.

ID	Recall (%)	Precision (%)
1	95,81	97,42
2	94,61	86,62
3	100,00	53,85
4	100,00	100,00
5	84,37	96,43
6	87,50	100,00

The second set of experiments evaluates the performance of the proposed character and character ligature recognition approach by measuring the classification performance of state-of-the-art classification algorithms. The experiments focus on characters and character ligatures that correspond to patterns with ID 1-2 (Table 1), since these patterns appear in a great variety of characters. Furthermore, patterns with ID 4-6 correspond to a unique character or character ligature and its subsequent detection implies a direct classification. The experiment involved two steps: the training and testing of a statistical classifier. The focus of the experiments was on testing suitability of the extracted features.

To that end, characters and character ligatures coming from 3 different writers (referenced as wr1, wr2 and wr3, respectively – Table 4) were gathered into two different datasets CL1 and CL2. CL1 dataset corresponds to Pattern ID 1, while CL2 dataset corresponds to Pattern ID 2 (Table 1). In Table 4, in the columns that concern the Training and Test set, we clearly indicate the percentage of the corresponding characters used in the experiment. More specifically, for the training stage, we use different percentages of the complete character set from writers wr1 and wr2, while for the Testing stage, we use the remaining percentages for the complete character set for either the case of writer wr1 and wr2 or the case of writer wr3. Within each dataset, the feature extraction

algorithm was applied to each character or character ligature. To measure the generalization performance of the trained classifiers, a splitting of data is necessary. Thus, a series of different scenarios with various ways of splitting has been constructed. Table 4 lists the scenarios for CL1 and CL2.

**Table 4.** Training set / Test set configuration for the CL1 and CL2 datasets

ID	Samples	Training	Test
CL1-1	754	70 (10% wr1, wr2)	684 (90% wr1, wr2)
CL1-2	754	147 (20% wr1, wr2)	607 (80% wr1, wr2)
CL1-3	479	147 (20% wr1, wr2)	332 (100% wr3)
CL1-4	402	70 (10% wr1, wr2)	332 (100% wr3)
CL1-5	1086	754 (100% wr1, wr2)	332 (100% wr3)
CL2-1	123	12 (10% wr1, wr2)	111 (90% wr1, wr2)
CL2-2	123	24 (20% wr1, wr2)	99 (80% wr1, wr2)
CL2-3	73	12 (10% wr1, wr2)	61 (100% wr3)
CL2-4	85	24 (20% wr1, wr2)	61 (100% wr3)
CL2-5	184	123 (100% wr1, wr2)	61 (100% wr3)

The classification step was performed using two well known classification algorithms, K-NN and Support Vector Machines (SVM) [Theodoridis, 1997]. K-NN was used in two variants, with L1 norm and L2 norm. Moreover, exhaustive search took place in order to determine the value of neighbors ( $k$ ) that gave the best score. On the other hand, SVM was used in conjunction with the Radial Basis Function (RBF) kernel, a popular, general-purpose yet powerful kernel. Again, a grid search was performed in order to find the optimal values for both the variance parameter ( $\gamma$ ) of the RBF kernel and the cost parameter ( $c$ ) of SVM. The results for CL1 and CL2 datasets along with the optimal parameter values are listed in Table 5 and Table 6, respectively. The scores that were achieved in both datasets were very high even in cases where the samples were few. This particular aspect is very encouraging, since it proves the good generalization performance of the algorithms. Furthermore, the fact that the algorithms were able to generalize so well, is also due to the robust feature extraction scheme.

**Table 5.** Algorithmic performance for the CL1 dataset. Numbers in parenthesis represent the parameters used for achieving the optimal scores. The ID column corresponds to the different scenarios as shown in Table 4. For the SVM kernel, the number of support vectors found is also given.

ID	KNN-L1	KNN-L2	SVM-rbf
CL1-1	90.49 ( $k=1$ )	90.78 ( $k=1$ )	<b>93.42</b> ( $\gamma=0.94$ , $c=50$ , SVs=45)
CL1-2	94.06 ( $k=1$ )	93.73 ( $k=1$ )	<b>95.05</b> ( $\gamma=0.98$ , $c=50$ , SVs=62)
CL1-3	<b>97.89</b> ( $k=2$ )	97.59 ( $k=2$ )	<b>97.89</b> ( $\gamma=0.04$ , $c=50$ , SVs=63)
CL1-4	94.27 ( $k=1$ )	96.08 ( $k=1$ )	<b>97.28</b> ( $\gamma=0.2$ , $c=100$ , SVs=42)
CL1-5	98.19 ( $k=10$ )	98.19 ( $k=4$ )	<b>98.49</b> ( $\gamma=0.8$ , $c=1$ , SVs=216)

**Table 6.** Algorithmic performance for the CL2 dataset. Numbers in parenthesis represent the parameters used for achieving the optimal scores. The ID column corresponds to the different scenarios as shown in Table 4.

ID	KNN-L1	KNN-L2	SVM-rbf
CL2-1	95.49 ( $k=1$ )	93.69 ( $k=1$ )	<b>94.59</b> ( $\gamma=0.1, c=10, SVs=9$ )
CL2-2	93.93 ( $k=1$ )	92.92 ( $k=1$ )	<b>94.94</b> ( $\gamma=0.1, c=20, SVs=10$ )
CL2-3	<b>100.0</b> ( $k=1$ )	<b>100.0</b> ( $k=1$ )	<b>100.0</b> ( $\gamma=0.1, c=10, SVs=9$ )
CL2-4	98.36 ( $k=6$ )	95.99 ( $k=1$ )	<b>100.0</b> ( $\gamma=0.1, c=10, SVs=11$ )
CL2-5	96.72 ( $k=1$ )	98.36 ( $k=1$ )	<b>100.0</b> ( $\gamma=0.1, c=10, SVs=24$ )

## Conclusion

D-SCRIBE project aims to support and facilitate current and future efforts in manuscript digitization and processing. In this paper, we focus on the problem of recognizing early Christian Greek manuscripts. We propose a novel digital image binarization scheme for low quality historical documents allowing further content exploitation in an efficient way. Additionally, we present a novel methodology that assists recognition of early Christian Greek manuscripts. We strive toward an assessment of the recognition procedure by tracing and recognizing the most frequently appearing characters or character ligatures, using a segmentation-free, quick and efficient approach. Based on the observation that closed cavities appear in the majority of characters and character ligatures, we propose a recognition technique that consists of several distinct stages. Experimental results show that the proposed method gives highly accurate results that offer a great assistance to old Greek handwritten manuscript interpretation.

Future work involves the detection and recognition of the remaining old Greek handwritten character and character ligatures that do not include closed cavities, as well as the testing of the performance of the proposed technique for other types of old handwritten historical manuscripts.

## Bibliography

- [Chen, 2003] C. H. Chen, J. de Curtins. Word Recognition in a Segmentation-Free Approach to OCR. Second International Conference on Document Analysis and Recognition (ICDAR'93), p. 573-576, 2003.
- [Duda, 1973] R. Duda, E. Hart. Pattern Classification and Scene Analysis, Wiley 1973.
- [Gatos, 2004] B. Gatos, I. Pratikakis, S. J. Perantonis. Locating Text in Historical Collection Manuscripts. Lecture Notes on AI, SETN 2004, p. 476-485, 2004.
- [Gonzalez, 1992] R. C. Gonzalez, R. E. Woods. Digital Image Processing, Addison-Wesley, 1992.
- [Hirano, 2001] T. Hirano, Y. Okada, F. Yoda. Field Extraction Method from Existing Forms Transmitted by Facsimile. Sixth International Conference on Document Analysis and Recognition, ICDAR2001, p. 738-742, 2001.
- [Jung, 1996] D. M. Jung, M. S. Krishnamoorthy, G. Nagy, A. Shapira. N-tuple features for OCR revisited, IEEE Trans. PAMI vol. 18, no. 7, p. 734-745, 1996.
- [Kavallieratou, 2002] E. Kavallieratou, N. Fakotakis, G. Kokkinakis. Handwritten character recognition based on structural characteristics. 16<sup>th</sup> International Conference on Pattern Recognition, p. 139-142, 2002.
- [Kim, 1996] I. K. Kim, R. H. Park. Local adaptive thresholding based on a water flow model. Second Japan-Korea Joint Workshop on Computer Vision, Japan, p. 21-27, 1996.
- [Lu, 2002] Y. Lu, C. L. Tan. Combination of multiple classifiers using probabilistic dictionary and its application to postcode recognition. Pattern Recognition 35, p. 2823-2832, 2002.
- [Mori, 1992] S. Mori, C. Y. Suen, K. Yamamoto. Historical review of OCR research and development, Proc. IEEE, vol. 80, p.1029-1058, 1992.
- [Niblack, 1986] W. Niblack. An Introduction to Digital Image Processing. Englewood Cliffs, N.J., Prentice Hall p.115-116, 1986.
- [Otsu, 1979] N. Otsu. A threshold selection method from gray-level histograms. IEEE Trans. Systems Man Cybernet. 9 (1) p.62-66, 1979.

- [Sauvola, 2000] J. Sauvola, M. Pietikainen. Adaptive Document Image Binarization. Pattern Recognition 33, p.225-236, 2000.
- [Suen, 1993] C. Y. Suen et al. Building a New Generation of Handwriting Recognition Systems. Pattern Recognition Letters 14, p. 303-315, 1993.
- [Theodoridis, 1997] S. Theodoridis, K. Koutroumbas. Pattern Recognition, Academic Press, 1997.
- [Vinciarelli, 2002] A. Vinciarelli . A survey on off-line Cursive Word Recognition. Pattern Recognition 35, p. 1433-1446, 2002.
- [Xia, 2003] F. Xia. Normal vector and winding number in 2D digital images with their application for hole detection. Pattern Recognition 36, p. 1383-1395, 2003.
- [Xu, 2001] Q. Xu, L. Lam, C. Y. Suen. A Knowledge-based Segmentation System for Handwritten Dates on Bank Cheques. Sixth International Conference on Document Analysis and Recognition, ICDAR2001, p. 384-388, 2001.

---

### Authors' Information

---

**Stavros J. Perantonis** – Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Research Center "Demokritos"; 153 10 Athens, Greece; e-mail: [sper@iit.demokritos.gr](mailto:sper@iit.demokritos.gr)

**Basilis Gatos** – Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Research Center "Demokritos"; 153 10 Athens, Greece; e-mail: [bgat@iit.demokritos.gr](mailto:bgat@iit.demokritos.gr)

**Konstantinos Ntzios** – Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Research Center "Demokritos"; 153 10 Athens, Greece; e-mail: [ntzios@iit.demokritos.gr](mailto:ntzios@iit.demokritos.gr)

**Ioannis Pratikakis** – Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Research Center "Demokritos"; 153 10 Athens, Greece; e-mail: [ipratika@iit.demokritos.gr](mailto:ipratika@iit.demokritos.gr)

**Athanasios Drigas** – Net Media Laboratory, National Research Center "Demokritos"; 153 10 Athens, Greece; e-mail: [dr@imm.demokritos.gr](mailto:dr@imm.demokritos.gr)

**Ioannis Vrettaros** – Net Media Laboratory, National Research Center "Demokritos"; 153 10 Athens, Greece; e-mail: [dr@imm.demokritos.gr](mailto:dr@imm.demokritos.gr)

**Christos Emmanouilidis** – ZENON S.A., Automation Technologies, R&D Division; Kanari 5, Glyka Nera Attikis 153 54, Athens, Greece; e-mail: [christosem@zenon.gr](mailto:christosem@zenon.gr)

**Anastasios Kesidis** – BSI S.A., R&D Division; 17 Noembriou 130, Xolargos 155 62, Athens, Greece; e-mail: [akes@bsi.gr](mailto:akes@bsi.gr)

**Dimitrios Kalomirakis** – Mount Sinai Foundation, Doryleou 26, Athens 121 15, Greece; e-mail: [kalomirakis@hotmail.com](mailto:kalomirakis@hotmail.com)

## MINERVA – THE MINISTERIAL NETWORK FOR VALORISING ACTIVITIES IN DIGITISATION TOWARDS AN AGREED EUROPEAN PLATFORM FOR DIGITISATION OF CULTURAL AND SCIENTIFIC HERITAGE

**Giuliana De Francesco**

*Abstract.* MINERVA is a project funded by the European Commission IST Programme within the 5th Framework Programme. It created a network of EU Ministries and other agencies in charge of cultural policies and programmes, which is open to enlargement to new countries and new sectors of the civil society. The network discusses, correlates and harmonises the activities carried out in the field of digitisation of cultural and scientific heritage, aiming at creating a common European platform made up of agreed recommendations, guidelines, standards. The network acts also to foster collaboration between European Commission and Member States, to ensure awareness of European policies at national level, to exchange good practice, to coordinate national programmes in order to embed in national digitisation activities the technical results achieved by the network. Some main outcomes of the activities are presented.



---

## Introduction: MINERVA Framework

---

**eEurope Initiative.** The twofold objective of the *eEurope Initiative: An Information Society for All*<sup>1</sup>, launched by European Commission in December 1999, was to bring the benefits of the information society to all European citizens and to make the European Union a most competitive and dynamic knowledge based economy, ensuring that the whole process was socially inclusive.

The objectives were to be achieved by supporting new infrastructure and services across Europe, applying the open method of coordination and benchmarking, and by accelerating the setting up of an appropriate legal environment. The need for political commitment was clearly stated as a condition to reach eEurope targets. eEurope 2002 Action Plan<sup>2</sup> focused on three main actions:

1. establishing a cheaper, faster, secure Internet
2. investing in people and skills
3. stimulating the use of the Internet. Objective 3d, encapsulated in the third action line, targeted the production of European digital content for global networks, in order fully to exploit the opportunities created by the digital technologies.

**Lund Principles.** eEurope 2002 Action Plan pointed out a specific action for Member States and the Commission jointly: to create a coordination mechanism for digitisation programmes across Member States. On 4th April 2001, representatives and experts of EU Member States met in Lund (Sweden) to agree on actions to support coordination and add value to digitisation activities in a sustainable way. Outcome of the meeting were the Lund Principles<sup>3</sup>. Starting from the statement that *Europe's cultural and scientific knowledge resources are a unique public asset forming the collective and evolving memory of our diverse societies and providing a solid basis for the development of our digital content industries in a sustainable knowledge society*, Lund Principles state that digitisation of heritage resources is a crucial activity for providing improved access for the citizen; for preserving Europe's collective cultural heritage (both past and future); for sustaining and promoting cultural diversity in a global environment. Digitised cultural assets are also a key resource for education and for the tourism and media industries. However, the full realisation of the potential of these resources is endangered by a number of key problems: among them, fragmentation of approach, obsolescence, lack of simple forms of access for the citizen, intellectual property rights management, lack of synergies between cultural and new technologies programmes and lack of institutional commitment. In order to address these issues, Member States had to establish an ongoing forum for coordination, to support the developing of a European view on policies and programmes, to exchange and promote good practices, guidelines and skills development, to work in a collaborative manner to make visible and accessible the digitised cultural heritage of Europe.

A set of actions for improving digitisation of cultural and scientific content in Europe and for achieving the objectives set out in the Lund Principles is described by the Lund Action Plan;<sup>4</sup> the set of actions is updated every year.

**The National Representatives Group (NRG).** The Lund meeting called upon Member States to establish a steering group for coordination of digitisation policies and programmes, with special emphasis on cultural and scientific resources. In 2001 the National Representatives Group<sup>5</sup> was established, made up of officially nominated representatives of all EU Member States, in order to monitor progress regarding the objectives encapsulated in the Lund Action Plan.

---

All urls quoted have been checked on October 18<sup>th</sup>, 2004.

<sup>1</sup> eEurope Initiative website: [http://europa.eu.int/information\\_society/eeurope/2005/index\\_en.htm](http://europa.eu.int/information_society/eeurope/2005/index_en.htm).

<sup>2</sup> eEurope 2002 Action Plan : [http://europa.eu.int/information\\_society/eeurope/2002/index\\_en.htm](http://europa.eu.int/information_society/eeurope/2002/index_en.htm).

It is worth mentioning that the current Action Plan is eEurope Action Plan 2005.

<sup>3</sup> The Lund Principles: [http://www.cordis.lu/ist/directorate\\_e/digicult/lund\\_principles.htm](http://www.cordis.lu/ist/directorate_e/digicult/lund_principles.htm).

<sup>4</sup> The Lund Action Plan: [http://www.cordis.lu/ist/ka3/digicult/lund\\_ap\\_browse.htm](http://www.cordis.lu/ist/ka3/digicult/lund_ap_browse.htm).

<sup>5</sup> The National Representatives Group: <http://www.MINERVAeurope.org/structure/nrg.htm>.

The objectives of the NRG, as described by the NRG Terms of reference<sup>1</sup>, are to share national experiences; to create a common platform for cooperation and coordination of national activities across the European Union; to provide the focus for consensus building among Member States for coordinating policies and programmes and for directing their follow up at national level; to promote good practice, skills development and training. The NRG identifies and nominates experts for the workgroups set up for the implementation of the Lund Action Plan, and validates reports, studies and recommendations prepared by the workgroups; each NRG member disseminates the results and promotes discussion on emerging issues in the own country.

The NRG meets every six months, under the chairmanship of the representative of the rotating EU Presidency, supported by a representative of the European Commission.

The *Charter of Parma*<sup>2</sup>, a strategic document that continues and reinforces the Lund principles, was approved by the NRG, gathered in Parma on the 19<sup>th</sup> November 2003 for their 5<sup>th</sup> official meeting. The Charter highlights progresses and outlines strategies, aiming also at consolidating the NRG position and high-level political commitment.

The 7<sup>th</sup> NRG meeting took place in The Hague on 17<sup>th</sup> September 2004<sup>3</sup>.

**The NRG Report.** An important task of the NRG is the production of a report illustrating progress in the coordination initiative and presenting an overview of each Member State's national policies, programmes and networks in the field of digitization of cultural and scientific heritage. *Coordinating Digitization in Europe: Progress Report of the National Representatives Group*<sup>4</sup>, edited by the MINERVA project on behalf of the European Commission, is simultaneously an outcome and a tool of the coordination initiative. In 2004 was published the second issue of the Progress report, referring to the policies, programmes and activities carried out in 2003.

---

### **MINERVA Project**

---

Funded by the European Commission IST<sup>5</sup> programme within the Fifth Framework Programme for Research and Technological Development, the Ministerial NETwork for Valorising Activities in digitisation (MINERVA) established a network of EU Member States Ministries in charge of cultural policy, aiming at creating a common European vision on policies and programmes and at harmonising the activities carried out in the field of scientific and cultural heritage digitisation.

MINERVA supports the activities of the NRG as an 'operative arm' and works to accomplish the Lund Action Plan.

The project is coordinated by the Italian Ministry for Cultural Heritage and Activities.

**MINERVA's twin-track action.** The attempt to coordinate digitisation policies and programmes, in order to avoid fragmentation and waste of resources, requires both political/institutional strategies and technical tools. Therefore MINERVA acts on both a 'political' and a technical level. The main political effort is made to guarantee closer collaboration among Member States and between these and the European Commission. The network gives international visibility to national initiatives, thus promoting the exchange of good practices, and ensures the awareness of EU policies and MINERVA achievements at national and local levels. The strategic impact of the network helps in several countries the coordination of national programmes, and stimulates the arising of several national digitisation programmes under the aegis of MINERVA.

---

<sup>1</sup> *Terms of Reference of the NRG:* [http://www.cordis.lu/ist/directorate\\_e/digicult/t\\_reference.htm](http://www.cordis.lu/ist/directorate_e/digicult/t_reference.htm).

<sup>2</sup> The *Charter of Parma* is available in English, French, German and Italian at: <http://www.MINERVAeurope.org/structure/nrg/documents/charterparma.htm>.

<sup>3</sup> The conclusions of NRG meetings are published on MINERVA website: <http://www.MINERVAeurope.org/structure/nrg/meetings.htm>.

<sup>4</sup> NRG Report is both printed and published on MINERVA website: <http://www.MINERVAeurope.org/publications/globalreport.htm>.

<sup>5</sup> Information Society Technologies (IST): <http://www.cordis.lu/ist/>.

The network is open, and aims at establishing contacts and cooperation with other countries, international and national organisations, associations, networks and projects engaged in the digitisation field.

MINERVA action at technical level concerns the workgroups for the implementation of the Lund Action Plan. They are further presented in detail.

**MINERVAplus.** The MINERVA project started in March 2002 with 7 partners; the remainder of the then EU member states joined the network during the first year of activity. The enlargement of the EU with ten new Member States in 2004 provided the opportunity to enlarge the network to the New Accession States (NAS). MINERVAplus<sup>1</sup> is the name of the project, financed in the framework of Sixth Framework Programme, which officially allowed the extension of MINERVA project to six NAS, plus Russia and Israel; the kick off meeting took place in Budapest in February 2004. MINERVAplus too is coordinated by the Italian Ministry for Cultural Heritage and Activities. MINERVA and MINERVAplus action lines are synchronised.<sup>2</sup>

**MINERVA Website.** A strong effort is devoted by MINERVA to the dissemination of results; its most important tool is the website. Every information, document or publication produced in the framework of the various Minerva action lines is published and/or uploaded on the website [www.minervaeurope.org](http://www.minervaeurope.org), which is constantly updated and provides other relevant information in the field of digitisation of cultural and scientific heritage. Minerva website is therefore considered as a knowledge base. Documents and publications are freely downloadable, for non commercial purposes only, and other websites are permitted to link them, even if they are not allowed to include MINERVA's contents without permission.

---

### MINERVA Workgroups

---

The real 'engine' of the network is represented by the thematic working groups. The goal of the workgroups is to establish a common European platform made up of shared recommendations, guidelines, metadata and standards related to digital cultural content creation and access; Another goal is to foster the convergence among archives, libraries, museums in a perspective of integration of the services offered by the memory institutions. MINERVA technical workgroups deal with interoperability and long-term preservation of resources, discovery of cultural content, benchmarking and good practice, quality and accessibility of cultural websites. The activity carried on in the framework of MINERVA workgroups by the new MINERVAplus partners is focused on the deeper investigation of four specific topics, as further explained under each working group.

**WP2, Benchmarking framework.** In line with the eEurope approach, Member States representatives have recognised the value of benchmarking, an ongoing search for best practices that produce superior performance when adapted and implemented in another organization, as a tool for exchanging experience and learning from good practice. The aim of MINERVA workpackage 2<sup>3</sup> was to outline key issues and to identify a method and a model for benchmarking digitisation programmes and projects, thus developing an innovative approach for the cultural heritage sector. Benchmarking activity closed by the end of 2003, delivering the final report *New Opportunities for Benchmarking the Digitisation of Cultural Heritage in Europe*.<sup>4</sup> Benchmarking is now considered an ongoing method crossing every workgroups' activity, and particularly that of WP6.

**WP6, Identification of good practices and competence centres.** Expertise and skills on digitisation are widely available across Europe, and a key issue is to feedback the experience and expertise developed within projects.

---

<sup>1</sup> Project presentation at: <http://www.MINERVAeurope.org/whatis/MINERVAplus.htm>.

<sup>2</sup> Some partners have set up MINERVA websites in national languages: Hungarian MINERVAplus website: <http://www.mek.oszk.hu/MINERVA/>; Israeli MINERVAplus website: <http://www.ejewish.info/reka/minerva/index.htm>; Russian MINERVAplus website: <http://www.minervaplus.ru/>.

<sup>3</sup> Benchmarking framework workgroup presentation: <http://www.minervaeurope.org/structure/workinggroups/benchmarking.htm>; Reports and documents related to the activity of the workgroup are to find at the following url: <http://www.MINERVAeurope.org/structure/workinggroups/benchmarking/docindex.htm>.

<sup>4</sup> Benchmarking full text report at: <http://www.minervaeurope.org/intranet/documents/benchmarkingreport2.pdf>.

WP6<sup>1</sup> deals with criteria for the selection of good practices, benchmarking as support for evaluation, promotion of guidelines and recommendations as extracted from good practice examples. Aim of the group is also to undertake the identification of specialised ("advisory" or "competence") centres at national and European level.

A first collection of good practices was carried out in 2002; A second campaign for the collection of good practices in digitisation and of information on competence centres started in 2004, in order to involve new MINERVAplus partners and to update MINERVA Knowledge Base with the collection of a critical mass of examples. The campaign is still ongoing, and both Good Practice and Competence Centre nomination forms are available on line,<sup>2</sup> so that interested organisations can submit the forms directly through the MINERVA website.

The main outcome of the activities carried out by MINERVA WP6, the *Good Practice Handbook*<sup>3</sup> is conceived as a practical guide for the establishment, execution and management of digitisation projects, with particular focus on the cultural area. The target audience of the handbook is teams within and across cultural institutions contemplating, or already executing, digitisation projects. The core text is articulated into ten categories, corresponding to as many steps of the digitisation process: Digitisation Project Planning; Selecting Source Material for Digitisation; Preparation for Digitisation; Handling of Originals; The Digitisation Process; Preservation of Digital Master Material; Meta-Data; Publication; IPR and Copyright; Managing Digitisation projects. The first printed edition of the Handbook was quickly exhausted. A second, more concise edition has just been published in English, French, German, Italian, Portuguese, and Slovak. The Handbook is to be complemented by the selected list of digitisation guidelines<sup>4</sup> and the MINERVA good practices list, published on the website.

**WP6 Specific Topic.** On the basis of the activity carried out and tools produced by the working group, the MINERVAplus WP6 will propose a model for digitisation cost reduction for the validation through experimental actions. A preliminary result of this activity is the survey *Good practices in cost reduction for digitisation: resources for Minerva and Minerva Plus WG on good practices.*<sup>5</sup>

**WP3, Inventories, discovery of digitised content, multilingualism issues.** This working group<sup>6</sup> deals with visibility and accessibility of European digital cultural and scientific content. The group addressed the definition of a sustainable technical infrastructure for coordinated discovery of European digitised cultural and scientific content, including a common set of metadata for the description of digital cultural collections; the analysis of possible solutions to make content accessible across different languages (multilinguality); the proposal of a common platform (XML and open source) for accessing distributed information in Europe.

Among the main outcomes is the *Specifications for Inventories of Digitised Content*<sup>7</sup>, already adopted by several national programmes and projects, and basis of MICHAEL project (see further).

**WP3 Specific Topic.** In the framework of MINERVA plus, WP3 is focusing the activity on the specific topic "Multilingual thesauri". The first stage is a survey<sup>8</sup> on the implementation of tools for multilingual retrieval by European cultural websites, which is particularly focused on controlled vocabularies and multilingual thesauri.

---

<sup>1</sup> Presentation of WP6 at: <http://www.minervaeurope.org/structure/workinggroups/goodpract.htm>. WP6 Reports and documents can be referred to at: <http://www.minervaeurope.org/structure/workinggroups/goodpract/docindex.htm>.

<sup>2</sup> The campaign for the collection of good practices is presented at: <http://www.MINERVAeurope.org/goodpractcamp.htm>; the campaign for the collection of information on competence centres at: <http://www.MINERVAeurope.org/competencentrecamp.htm>.

<sup>3</sup> Good Practice Handbook: <http://www.MINERVAeurope.org/publications/goodhand.htm>.

<sup>4</sup> Digitisation guidelines: a selected list (<http://www.MINERVAeurope.org/guidelines.htm>).

<sup>5</sup> The study is to read at:

<http://www.minervaeurope.org/structure/workinggroups/goodpract/costreduction/documents/wp6costreduction0904.pdf>.

<sup>6</sup> Workgroup presentation: <http://www.minervaeurope.org/structure/workinggroups/inventor.htm>. Reports and documents related to the activity of WP3 are to find at the following url:

<http://www.minervaeurope.org/structure/workinggroups/inventor/docindex.htm>

<sup>7</sup> The *Specifications* is published just online (<http://www.minervaeurope.org/intranet/documents/specinv0311.pdf>).

<sup>8</sup> Survey of Multilingualism and the Use of Controlled Vocabularies of Cultural Sites in MINERVA Countries:

<http://www.mek.oszk.hu/MINERVA/survey/survey.html>

**WP4, Interoperability and service provision.** This action line was set up with the ambitious objective of identifying a European common framework for an information environment allowing the delivery of shared services and the integrated access to digital cultural resources.<sup>1</sup> The goal was to be achieved through analysis and comparison of international and national approaches, activities, research and best practice concerning technical and metadata standards. Main outcome of this action line is the *Technical Guidelines for Digital Cultural Content Creation Programmes* (presented further). **WP4 subgroup IPR, copyright and data protection.** In some countries a specific sub-group was set up to address IPR issues,<sup>2</sup> because of their wideness and complexity. This sub-group deals with how rights can be assigned to or shared with funding bodies, negotiated with licensing agencies, special provision be made for free access for educational or other specific uses, how cultural institutions can exploit commercial rights for reuse, as well as other legal and regulatory issues, such as privacy, data protection, freedom of information, security.

**WP4 Specific Topic.** The study of IPR issues is particularly focused in the framework of MINERVAplus. The University of Athens, Greek partner of the network, continues the activity undertaken by English and Italian working groups, aiming at defining a business model for digitisation projects which will represent a specific tool for the management of IPR and related issues.

**WP5, Identification of users' needs, content and quality framework for common access points.** The objective of this workgroup is to provide a shared vision of quality criteria for websites intended to give access to cultural and scientific contents, and to encourage the use of accessibility and quality framework in cultural websites, thus facilitating the networking of cultural information.<sup>3</sup>

The workgroup published the *Handbook for Quality in Cultural Websites: Improving Quality for Citizens*;<sup>4</sup> it arises from the Brussels Quality Framework, edited in 2001,<sup>5</sup> which proposes some criteria for accessibility and quality of cultural websites. An English version and an Italian edition of the Handbook have been issued, a German version is foreseen.<sup>6</sup> The Italian WP5 is at present checking the criteria proposed by the handbook through a programme of test beds.<sup>7</sup> The results of the assessment will be presented in Spring 2005.

The basic concepts of the quality handbook have been condensed into ten *Cultural Website Quality Principles*, whose translation into several European languages is available on MINERVA website, but has also been disseminated through printed posters and postcards.<sup>8</sup> The European WP5 has further developed an explanatory document on the criteria for the adoption of the ten Principles.<sup>9</sup>

**WP5 Specific topic.** The specific objective of MINERVAplus WP5 activity is to provide concrete tools for the creation and management of quality cultural content for the Web to small and medium cultural institutions, adapting and disseminating the MINERVA WP5 products.

---

<sup>1</sup> Workgroup presentation at: <http://www.minervaeurope.org/structure/workinggroups/servprov.htm>. Reports and documents produced by WP4 can be referred to at: <http://www.minervaeurope.org/structure/workinggroups/servprov/docindex.htm>

<sup>2</sup> IPR Subgroup presentation: <http://www.minervaeurope.org/structure/workinggroups/servprov/ipr.htm>

<sup>3</sup> Presentation of the workpackage at: <http://www.minervaeurope.org/structure/workinggroups/userneeds.htm>. Reports and documents are to find at: <http://www.minervaeurope.org/structure/workinggroups/userneeds/docindex.htm>.

<sup>4</sup> Handbook webpage: <http://www.MINERVAeurope.org/publications/qualitycriteria.htm>.

<sup>5</sup> Brussels Quality Framework website: <http://www.cfwb.be/qualite-bruxelles/>.

<sup>6</sup> *Manuale per la qualità dei siti Web pubblici culturali*: <http://www.MINERVAeurope.org/publications/qualitycriteria-i.htm>.

<sup>7</sup> For information on Handbook testing:

<http://www.MINERVAeurope.org/structure/workinggroups/userneeds/handbooktest-i.htm>.

<sup>8</sup> Cultural Website Quality Principles: <http://www.MINERVAeurope.org/structure/workinggroups/userneeds/documents/cwqp-uk.htm> The text of the principles has been so far translated into: English, French, Spanish, German, Greek, Estonian, Hungarian, Italian, Slovenian.

<sup>9</sup> Commentary and Exploration of the Ten Quality Principles

[http://www.minervaeurope.org/publications/qualitycommentary\\_en.htm](http://www.minervaeurope.org/publications/qualitycommentary_en.htm). Also available in Hungarian.

---

## Digitisation Cluster

---

MINERVA promoted the set up of a so-called cluster of the European networks and projects dealing with digitisation of culture heritage. Purpose of the initiative was to promote reciprocal cooperation among the networks in order to avoid duplication of efforts, maximise the effectiveness of the projects and define a common research area in the field of digitisation of cultural heritage.<sup>1</sup>

The digitisation cluster met for the first time in Rome in October 2003, on which occasion a first set of issues to tackle was agreed.

The cluster gathers at present the following projects, mostly funded by IST 5<sup>th</sup> and 6<sup>th</sup> FP: BRICKS, CALIMERA, DELOS, DIGICULT, EPOCH, EMII-DCF, ERPANET, EVA network, HEREIN, MUSICNETWORK, PRESTOSPACE, EUROMED HERITAGE II and SCRAN.

Among the outcomes of such coordination initiatives:

- MINERVA and EMII-CDF joint position paper *Encouraging IST research on European digital cultural content, fostering a common strategic action line to enable the transfer of new technologies to memory organisations*<sup>2</sup>.
- MINERVA and ERPANET workgroup on long term preservation of digital memories.

---

## Technical Guidelines for Digital Cultural Content Creation Programmes

---

MINERVA WP4 promoted the cooperation among IST-funded cultural networks on technical and metadata standards. The main outcome, the *Technical Guidelines for Digital Cultural Content Creation Programmes*,<sup>3</sup> was drafted in cooperation with the PULMAN, EMII-DCF and ERPANET projects and proposed for adoption and improvement to all cluster projects.

Throughout Europe, international, national, regional and local initiatives are investing significant public and private sector funding to enable access to a range of cultural heritage resources through digital channels. In order that the content produced is as widely useful, portable and durable as possible, resources should be interoperable, accessible, preservable and secure. *Technical Guidelines* suggest the appropriate use of standards in digitisation as the way to ensure for digital resources the consistency that makes interoperability possible.

Intended primarily as a resource for policy-makers, and for those implementing funding programmes for the creation of digital cultural content, *Technical Guidelines* is not intended to be a single prescriptive set of requirements to which all projects must conform, but seeks to identify those areas in which there is already a commonality of approach and to provide a core around which context-specific requirements might be built. TG structure reflects a 'life cycle' approach to the digitisation process, paralleled in MINERVA *Good Practice Handbook*, which emphasises the importance of seeing the project as a whole, and how decisions taken at given stages have implications for the rest of the process and affect the continuing development of the service. The guidelines text distinguishes between requirements and guidance using the keywords 'must' 'should' and 'may', based on Internet Engineering Task Force (IETF) terminology. Within each section guidance on practice and detailed standards is provided, and links are provided to sources of further guidance and information when available.

*Technical Guidelines* aims at representing a firm foundation for the development of interoperable trans-European services and will be maintained and updated. To foster the adoption by national digitisation programmes and by European projects dealing with digitisation of cultural heritage, *Technical Guidelines* are being translated,

---

<sup>1</sup> Digitisation cluster webpage: <http://www.MINERVAeurope.org/digiclust.htm>.

<sup>2</sup> <http://www.MINERVAeurope.org/intranet/documents/emiiMINERVA0310.pdf>.

<sup>3</sup> Technical Guidelines webpage: <http://www.MINERVAeurope.org/publications/technicalguidelines.htm>; full text of the English version 1.0 at: [http://www.minervaeurope.org/structure/workinggroups/servprov/documents/techguid1\\_0.pdf](http://www.minervaeurope.org/structure/workinggroups/servprov/documents/techguid1_0.pdf).



---

updated and localised. To date, a French translation including French references is already available on the website.<sup>1</sup>

An Italian workgroup has been established to issue an Italian edition, taking into account national standards and guidelines, to support the creation of digital resources in the framework of the forthcoming National Multilingual Portal of the Cultural and Tourism Resources.

---

### **MINERVA Spin Offs: MICHAEL**

---

MICHAEL is the first MINERVA spin-off, based on the joint efforts of Italy, France and United Kingdom on interoperability and inventories carried out within MINERVA, and was approved and funded in the framework of eTen programme, based on national funding.

The project acronym stands for Multilingual Inventory of Cultural Heritage in Europe;<sup>2</sup> it aims at establishing a trans-European inventory of digital cultural collections and resources and an international online service allowing the users to search, browse and examine them, also enabling the search across multiple national cultural portals from a single access point.

The MICHAEL consortium is made up of the French Ministère de la culture et de la communication, the UK Museums, Archives and Library Council and the Italian Ministero per i beni e le attività culturali, which is the project coordinator. The Ministries are supported by three private partners for the technological and administrative aspects.

Participation of other European countries is expected; some candidate partners have already expressed their interest, and will start their activity once a stable MICHAEL instance will be ready (first half of 2005).

MICHAEL platform is based on standards and open-source technologies that allow flexibility and extensibility, and builds upon the following existing assets:

- the technical platform used in the French *Catalogue des fonds culturels numérisés*;<sup>3</sup>
- the metadata model for inventories of digital collections developed by MINERVA WP3;
- the French-Italian Prototype Portal, developed by MINERVA WP3 based on the two previous assets;
- European standards, methodology and further guidelines established by MINERVA and agreed and validated by the NRG.

The end user will exploit MICHAEL services to find and explore European digital cultural heritage, made accessible over the Internet on a multilingual basis. MICHAEL will thus make further progress towards the implementation of Lund Action Plan.

---

### **Author Information**

---

**Giuliana De Francesco** – Ministry for Cultural Heritage and Activities; Directorate General for Innovation and Promotion; Via del Collegio Romano, 27; 00185 Rome; Italy; e-mail: [defrancesco@beniculturali.it](mailto:defrancesco@beniculturali.it)

---

<sup>1</sup> Recommandations techniques pour les programmes de création de contenus culturels numériques:  
[http://www.minervaeurope.org/structure/workinggroups/servprov/documents/techguid1\\_0-f.pdf](http://www.minervaeurope.org/structure/workinggroups/servprov/documents/techguid1_0-f.pdf).

<sup>2</sup> MICHAEL project website: [www.michael-culture.org](http://www.michael-culture.org).

<sup>3</sup> [http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f\\_02.htm](http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm).

## DML AND RUSDML – VIRTUAL LIBRARY INITIATIVES FOR COVERING ALL MATHEMATICS ELECTRONICALLY

**Bernd Wegner**

*Abstract: With the rapidly growing activities in electronic publishing ideas came up to install global repositories which deal with three mainstreams in this enterprise: storing the electronic material currently available, pursuing projects to solve the archiving problem for this material with the ambition to preserve the content in readable form for future generations, and to capture the printed literature in digital versions providing good access and search facilities for the readers. Long-term availability of published research articles in mathematics and easy access to them is a strong need for researchers working with mathematics. Hence in this domain some pioneering projects have been established addressing the above mentioned problems.*

*Keywords: DLM, EMANI, ERAM, RusDML*

---

### Introduction

---

The paper will give a short state of the art report on some of these activities on the European level and the world wide plan to develop a global Digital Library in Mathematics (DLM). For example, in the archiving area as a special project for mathematics the Electronic Mathematics Archives Network Initiative (EMANI) had been designed. Having in mind that a distributed architecture would be more suitable and reduce the load on the partners for such a project, a network is proposed, which also might be a more open approach for extending the project from a initially restricted solution to a more comprehensive enterprise.

The Electronic Research Archive in Mathematics (ERAM) is a DFG-funded German project dealing with capturing the content of a classical bibliographic service in mathematics in a database, and combining this with the retro-digitisation of selected mathematical publications. This is extended now by further projects which shall try to retro-digitise the national mathematical heritage in several countries world-wide. In particular ideas to cover the Russian publications in a digital repository called RusDLM have been implemented in a project funded by DFG and RFBR. As further digitisation projects the French activity NUMDAM, pursued by Cellule MathDoc in Grenoble, and the European Cooperation in DIEPER have to be mentioned.

---

### 1. Electronic Offers and Their Providers

---

The impact of electronic devices on the daily life of researchers, teachers or other professionals results from a variety of tools and offers installed in local machines or made accessible through the internet. The part libraries are mostly involved in consists of electronic publications, or better electronic versions of printed publications. Some libraries already developed digital repositories containing retro-digitised publications, which had been obtained by scanning printed articles and books. But also offers which could be published only in electronic form become more and more important. In addition to this researchers and teachers increasingly take advantage of computer algebra systems and other computing software, and visualisation techniques using graphics software and image processing tools have become background for most of their presentations and publications. Finally, we should not forget that the internet has been used to establish a communication infrastructure which strongly facilitates their daily work and extend the possibilities for co-operation at distributed sites.

There is a wide range of providers of these offers, going from commercial publishers and learned societies to volunteers and single authors. Also the list of distributors and information brokers is a long one: libraries, databases and indexing services, internet-portals of different types, web browsers et al. In contrast to the "old world" of printed publications these providers have different aims and it is not always clear for the user what he really could expect from these services, when he is searching for some information or article of his own interest. Clearly, libraries try to transfer their system, they have developed for their printed holdings, to these new publications, and hence they still seem to be the most reliable information provider also with respect to electronic offers. But this role has to be acknowledged more widely and the offer has to be improved.



There are good reasons why libraries will be able to maintain their central role for distribution and storage of scientific information and succeed to extend this to the electronic media. They have developed precise and reliable access structures. Their service is free for their specific group of users, and this group is a large one in most cases. Even for external users they developed a good network of exchange facilities, which enables scientists to make their work really accessible for a wide community of users and to read the work of their colleagues without being confronted with bigger commercial barriers. Commonly libraries cover a broad area of subjects and within that they try to be relatively comprehensive. Independent from the frequency of their usage these holdings had been preserved and kept accessible with great care. The objectives of science libraries are user-oriented on one side, and on the other side libraries feel obliged to protect the treasure of knowledge they have accumulated in their collection. This makes them also the best choice for solving the problem of the long-term preservation of electronic publications.

Mathematics is a science where the availability of electronic publications and retro-digitised documents lead to a considerable improvements of the conditions for research. Hence, though some of the subsequent arguments may apply to all sciences, they turn out to be of particular importance for mathematics: Mathematicians and professionals applying mathematics need quick, reliable and integrated access to mathematical publications. Long-term availability of publications is a particular need in mathematics. Digitising of print-only publications and the adjustment of these offers to the current facilities provided for electronic publications leads to a additional series of problems to be solved. Electronic publishing offers a variety of additional information in mathematics which may be integrated into the access and display structures enhancing the traditional types of publications.

---

## 2. Some Evidence by Numbers

---

For non-mathematicians it is not clear at all that mathematics is so much different from other sciences as far as easy availability of older publications will be concerned. For some it is even hard to understand the subjects of mathematical research and the special way how this research is published. For example, extensions and improvements of older results only care about the publication of the additional achievements, and there detailed proofs are essential. Older results may and should be cited, but it is not honest to repeat their proofs in research publications, even if the understanding of these proofs is essential for seeing what the new results are about. Many proofs can be found at one place only. Hence an article is just an addition to a sequence of other articles, more or less tightly interrelated in a structure which combinatorically is more complex than a tree. It provides another shell to a core of theorems, propositions, examples, models and proofs representing the current knowledge of a subject domain in mathematics. Mathematical research articles commonly are rather thin, and the publication frequency of a mathematician is rather low compared to other sciences.

Admittedly, parts of such a domain may be exhibited comprehensively in monographs, but as can be seen by the variety of material in the research surveys in mathematics which have been published by VINITI (Moscow), for example, such monographs with detailed exhibitions of arguments only can cover a part of the domain of reference, giving a motivating introduction with proofs, while the surveys have no space to provide proofs at all, if they really want to be comprehensive. This underlines that references in mathematical papers are not just a matter of honesty, but that at least a part of them plays an important role for a complete understanding of the content of an article. Hence the following figures give a good evidence for the need to have also older mathematical publications available.

The evidence will be demonstrated in the case of three journals where the numbers are taken from an investigation by Joachim Heinze [see *Joachim Heinze*]. The most surprising figures (also to mathematicians) are the numbers of citations before 1992. In the case of the most traditional mathematical journal from North-America, the *Annals of Mathematics*, 60% of the citations in the 35 articles published in that journal in 2001 had a publication date before 1992. Vice versa, the number of cites from the volumes of 500 journals published in 2001 to the *Annals* was about 4.500 and 82% of them were before 1992. Looking at one of the first journals which published mathematics only (in contrast to journals which deal with several sciences), the *Journal fuer die Reine und Angewandte Mathematik*, founded as *Crelles Journal* in 1826, the first figure was 61% and the second 65%. Finally, these numbers still were high for a more "modern" journal which had been founded in the second half of the 20th century, the *Inventiones Mathematicae*: the first figure was 55% and the second one 68%. Such high

numbers of older citations are not common for most of the other sciences. It would be quite interesting to have a more comprehensive comparison of this type.

---

### 3. Current and Future Problems

---

In the "paper world" the long-term preservation of publications was simple on the first view, though at a closer look a lot of problems had to be handled. They mainly came from the deterioration of the paper or the binding of a book or journal, and they appeared after a comparatively long period in which the physical situation of the document could be considered as stable. Also a wide distribution of documents to several locations world-wide was a factor of stability, protecting them against being all destroyed simultaneously by the impact of wars etc.

For digital publications this period of stability turned out to be extremely small. What everybody experiences with his old releases of word-files, became true meanwhile for the readers of PDF-files, for example. Without conversions, if they exist at all, or simultaneous installation of several versions of the Acrobat-reader a whole range of PDF-files over the period, where the Acrobat reader was offered, was not readable anymore recently. This admittedly was a temporary problem, because the next release of the Acrobat-reader was capable to handle the full range of previous productions. But nobody can guarantee that a similar bug will occur again in the future.

This is only one problem. Another one is the stability of the physical carrier, where the data are stored, and there is a variety of plug-ins which depend on additional software to be offered with the electronic document. Current releases of this software may have a short life-time. What should we do with the document afterwards?

To solve this problem will be even more complicated when documents in mathematics are considered, because they are most likely to have software depending enhancements. Interactive documents will play an important role in the future. Furthermore, projects like MoWGLI [A. Asperti; B. Wegner] will develop different types of structures enabling semantic mark-up of documents. Hence preservation will go far beyond caring about the displayed text only. Structures, links and other informational background provided with electronic articles will have to be taken care of, and all these tools are in permanent evolution.

Hence the problem only can be attacked by a long-term approach as it is described with in EMANI in the next section.

---

### 4. The EMANI Project

---

There is a period of approximately 10 years during which electronic publications in mathematics developed from some offers in pioneering freely accessible journals to a first class publication facility with enhanced services in comparison to traditional printed publications. As mentioned above, older publications are still of big value for research in mathematics. Hence retrospective digitisation projects increased the current digital content in mathematics considerably. One major of these projects is ERAM (see [H. Becker, B.Wegner] or [B.Wegner, ERAM]) which will be described later on.

In the first half of 2001, the Electronic Mathematics Archives Network Initiative (EMANI) had been founded as a special project to develop models for the archiving of electronic contents in mathematics. Having in mind that a distributed architecture would be more suitable and reduce the load on the partners for such a project, a network is proposed, which also might be a more open approach for extending the project from a initial restricted solution to a more comprehensive enterprise. The initiative has been formalised in July 2002 at their workshop at Cornell University with the partners mentioned below as the first set of members and the author of this article as the co-ordinator of the project.

Thus, for the core of the network, a co-operational system of reference libraries and content providers like publishers and editors has been set up. In the ideal final version they are supposed to serve for a long list purposes: The basic action will be to store the digital content in mathematics from the content providers at the reference libraries. This will be complemented by retro-digitising all printed publications in mathematics from the content providers at the reference libraries, covering a big part of the publications in mathematics by electronic versions finally. On this basis first measures can be undertaken to care about the long-term preservation of this content in readable form. First projects for the technical support of this co-operation have been just initiated.

For example, to have the content stored somewhere will not be sufficient. Retrospective digitisation may lead to scanned images only, which hopefully can be accessed in some repository. As an important enhancement it will be necessary to improve the usability of the retro-digitised publications by introducing advanced linking and

searching facilities and to provide convenient and affordable access to the stored content for mathematicians and professionals using mathematics world-wide.

The reference libraries even may serve as a reference system for other libraries which want to store and provide part of the content or refresh their existing offers by updated material. Having in mind the long time scale of the publications provided through the network, going from articles from the 19th century to current publications, a system of distribution agents will be needed. This may be a good reason to develop new business models for a distribution of mathematical publications in a combined enterprise between reference libraries and content providers. But there is not only a theoretical discussion about potential activities in the future.

---

## 5. The Starting Point of EMANI

---

It will be reasonable to start with such a complicated enterprise only on a smaller well-controllable scale at first. Once the architecture and the action plan will have been made sufficiently precise, an extension may be considered. The current partners who collaborate for the first steps in order to implement the initiative on the side of the libraries are:

- The Cornell University Library, Ithaca, N.Y.: They have a good tradition in retrospective digitisation projects and are involved in the archiving discussion for other sciences also. In particular they are building up an offer of a bundle of electronic journals in mathematics through project Euclid. They serve as a mirror site for EMIS (see [Bernd Wegner, ELibM]).
- The State and University Library Goettingen: Also there some important retrospective digitisation projects like ERAM (see [H. Becker, B.Wegner] or [B.Wegner, ERAM]) and DIEPER are pursued. In addition to this the SUB Goettingen is obliged to collect all publications in mathematics. In this role they have a high reputation as a centre for access to mathematical publications. Moreover they also serve as a mirror site for EMIS.
- The Tsinghua University Library, Beijing: This library has experience with the digitisation of Chinese publications. They are a Chinese centre of excellence for installing and offering electronic publications.
- The Orsay Mathematical Library, Paris, in co-operation with the Cellule MathDoc in Grenoble: The group in Orsay is co-ordinating a quite comprehensive consortium of French mathematical libraries. The strength of the partner in Grenoble consists in their excellent retro-digitisation project NUMDAM ([P. Berard]).

The content providers for the start are the publishers running under the group lead by Springer-SBM currently and the electronic library ELibM offered through EMIS, the European Mathematical Information Service (<http://www.emis.de>). At the beginning of 2004 Springer-SBM started to digitise all periodicals published by this group, independent from the subject. But the pioneering offer of several of the best journals in mathematics in EMANI could be preserved. In contrast to this the ElibM is a co-operation of several journals and editors on a voluntary basis bundling electronic offers in a world-wide system of WWW-servers (see [B. Wegner, ERAM]). They agree to provide open access in general.

An important step in the first phase of the initiative consists of the stepwise transfer of the available electronic content from the content providers to the reference libraries. There it will be checked if the files still can be used for the archiving, adjustments will be made in the case of files which are unsuitable for this and recommendations will be developed how the content providers could care about a more convenient delivery in future cases. Also new archiving related meta-data have to be defined, and an integrated access structure satisfying the needs of all kind of experts who want to work with the archive will be one of the central achievements of the further work in the future. Though links from reference databases could satisfy many of the needs of the mathematicians to get access, the professional handling of the archives will require more than just their meta-data.

To check the TEX-files for their usability without any appropriate system in the background turned out to be a tedious task. This was never considered as a part of the work of librarians, if we think of the period where TEX only was considered as a tool for preparing beautiful camera-ready manuscripts. But taking TEX as a tool for the mark-up of publications on the ASCII-level and providing files, which are most suitable for long-term preservation, the EMANI-partner Goettingen reacted very quickly, by establishing a project to develop tools for an automatic checking of the usability of TEX-files. A prototype for the tool could be presented in 2004.

## 6. ERAM – Combining a Database with an Archive

---

Also for older documents searchability will be an important requirement to enable the researcher to find his way in the huge knowledge base of mathematical achievements. Admittedly, no current search engine is able to locate a statement in its abstract meaning. Names for some of them will help, and classification codes of special subject areas will restrict the set of documents where to look for the desired information considerably. Hence literature databases for the classical period of mathematics are desirable. They should offer the same facilities like the current literature information services in mathematics, and even more, they should also provide links to the future given by modern mathematics. This is the starting point for the project ERAM which also will be called the Jahrbuch-project for short.

The acronym ERAM stands for "Electronic Research Archive for Mathematics". The project is funded by the Deutsche Forschungsgemeinschaft (DFG). The aim of the project is the installation of a (digital) archive of articles relevant for mathematical research, full searchability and access through a database, captured from the "Jahrbuch ueber die Fortschritte der Mathematik" (1868-1943). The most comprehensive current literature database in mathematics, Zentralblatt MATH, was founded at the end of the Jahrbuch period. The first step of the ERAM-project is the production of a bibliographic database, the JFM-database, capturing the content of the Jahrbuch ueber die Fortschritte der Mathematik (JFM). This has been finalized in a first version in the first half of 2003. Modern literature databases provide several search options for which the information could not easily be extracted from the text of the JFM. Hence, editorial enhancements are under preparation, and moreover historical links are provided to modern research as far as possible. The only formalised subject information in the JFM consists of the subject headings which are stored in the database like a raw classification. A more precise description of their subjects will be obtained by additional intellectual indexing work. The corresponding experts provide an English translation of the title of the single document, they add a subject classification according to the MSC2000 scheme and assign some English keywords.

All data from the Jahrbuch have been keyboarded now. They are made accessible in this form in the web, and though for many items the enhancements are still missing the database has found a lot of grateful users. As a consequence combined searches with the database Zentralblatt MATH are offered. In addition to its usage as high-quality source for information on classical mathematics, the JFM-database will provide access to a digital archive to be built up within the project. For this selected publications are scanned (as gif-images) and stored in a document management system. Currently there are no conversions of the images into text files. To allow text searches in the archive, text files will be an important addition to the scanned images. But the generation of these data will be a matter of a later phase of the project. Conversion programmes have been improved considerably, as had been demonstrated at a satellite meeting to the ECM 4 in Stockholm, and they are able to tackle the problems which occur with formulas in mathematical texts. A first step in this direction is made by a project based on the co-operation of experts from Japan, Germany and the United States (see [G. Michler]). But this has been topped by a conversion method provided by Tim Dokchitser meanwhile.

The scanned material includes journals like *Mathematische Annalen*, *Mathematische Zeitschrift*, *Inventiones Mathematicae*, *Commentarii Mathematici Helvetici*, for example. The *Journal fuer die Reine und Angewandte Mathematik* will be added at the end of 2004. Most of the journals which have installed recent electronic versions in EMIS (European Mathematical Information Service) agreed that all of their print-only back volumes could be digitised and offered within ERAM, and this also has been done. In ERAM, about 1 million pages have been scanned so far, and the capacity of the project will be sufficient to go for about 1.2 million pages. For more details see the references [H. Becker, B.Wegner] and [B.Wegner, ERAM], or the ERAM-homepage under <http://www.emis.de/projects/> clicking on the box for the Jahrbuch.

## 7. The Global Digital Mathematics Library – DML

---

ERAM could be considered as a part of a global initiative to have all mathematics digitally available. It has a lot of overlap with EMANI and both projects are tightly linked with each other. But in contrast to EMANI the global initiative at first will concentrate on retro-digitization, i.e. the preparation of digital versions of texts which are not yet digitally available. Long-term preservation is a secondary aspect of the DML at present. Clearly, in addition to ERAM there are several other digitization projects on the way, general projects like JSTOR, DIEPER, and the Elsevier backfiles system, and projects in mathematics like NUMDAM [P. Berard] or the national heritage activity

in Colombia by Victo Albis [V. *Albis*]. The Tsinghua University Library succeeded to digitise more than 50 Chinese journals. The digital offers made accessible through the EMANI homepage comprise more than 100 journals. At KISTI in Korea 16 mathematical journals have been digitised. All this has to be taken into account for getting an impression about the state of the DML.

In 2001 John Ewing prepared his White Paper [J. *Ewing*] in which a rough estimate has been made how much money would be needed to develop global digital mathematics library (DML) containing all mathematics in digital form. This estimate was in the order of 100 million US Dollars. But that was not the main achievement of that paper. It contained a lot of structural considerations for such a library, and it also addressed the immense problems we will be confronted with when we really want to pursue such a project. As a caveat when reading this paper, one should be aware that it describes an ideal solution, and some parts like a central repository (by intention) do not reflect very well what has been developed already. For example, at present only a system of distributed repositories could be imagined, because proprieties and aspects of cultural heritage have to be respected. Furthermore, a distributed system can hook on existing providers like libraries, and this will be more efficient than the installation of an extra infrastructure to manage the DLM, as far as the costs will be concerned.

As a consequence a planning grant had been given to Cornell University by the National Science Foundation of the U.S.A. to make a feasibility study for the DML. This will be done during two workshops, where the first one took place already in Washington D.C. at the end of July 2002, the second one at the SUB Goettingen in May 2003. The 25 participants from different kinds of institutions set up a scheme to develop a plan for the DML. An initiative has been formalised, working groups have been designed and a Steering Committee has been chosen to guide the progress of the discussion during the next future. More or less the scheme reflects a part of the project administration for EMANI, and indeed DML may profit a lot from the preparations in EMANI.

It will be the subject of an article of its own to go into all the details having been addressed by the working groups, but one of them should be explained here, because it is basic for the definition of the global project as well as for the description of the environment for similar national or local projects.

How can we determine what has to be considered as the content of mathematics?

Talking to mathematicians it will be noticed rather soon that the idea what should be covered by the DLM is quite vague. There are ongoing projects which have selected items for retrodigitization according to different aspects. These patches of the global DLM can be defined easily, but they cannot serve as a model for a comprehensive coverage of mathematical publications. Hence some more concrete questions arise naturally:

- Do we really have the chance and the interest to cover all mathematical publications world-wide by the DLM? If not so, the selection criteria have to be discussed. But also in the other case we have to decide on selection criteria, because not everything could be done immediately and a time schedule for building up the DLM step by step requires an order and hence a selection.
- Which publications may be considered as a part of mathematics according to subject area?

It will need a lot of efforts and patience to arrange such a list of contents and somebody has to administrate this. People are most likely to escape from this by deciding not to care about such a list at all and digitise what will be just in their mind or easily available. This is good for the patchwork, but it will ruin the global idea. To work on the global solution, four dimensions have to be considered: When do we start and how far back should we go? How much mathematics is supposed to be in the document? There are impact on research, potential user interest in having the document available, depending also on different user communities (research, education, applications, history etc.), quality, availability. Where should be the priorities? There is a geographical dimension which may be associated with priorities for the DLM-actions. How should the DLM project spread out from current initial activities covering content from all over the world ?

But there is also a cultural dimension. Though the global approach is a challenging idea, the development of the repositories should take national interests and funding possibilities into account. Hence distributing the content to single projects has to respect what had been covered already and what should remain under the guidance of a special mathematical community. Only the remaining content may be open for adoption for retro-digitisation. To distinguish this will be one of the main tasks during the content determination and it will be a delicate task, because very quickly there may be the impression that one party wants to buy out the mathematical heritage from another one.

The activities of the Planning Group have been finalized after the second meeting in Goettingen. The result will be a report to be presented to the NSF and to be made public by the chief researchers of this group. To go on with the coordination activities a committee has been installed by the IMU. But this is only one option to keep the DML initiative going. The DML-EU application is another integrating activity, which keeps the different parties talking to each other and promotes to think about the installation of digitisation projects being funded by other parties. One reason for this is that funding from the FP6 only can be used for networking and research activities. Basic digitisation only can be funded in the form of small test beds. Mass digitisation will have to ask for funding from other sources.

---

## 8. Aims and Mission of RusDML

---

The initial goal of RusDML (Russian Digital Mathematics Library) is to digitise a core collection of Russian journals in mathematics, which so far have been available in printed form only and, by making them accessible in the web, to facilitate the world wide access to them. Having succeeded with this a further activity may go for comprehensiveness, i.e., to perform the digitisation of all Russian mathematical publications, including monographs, series of collections of papers, encyclopaedias, handbooks, proceedings volumes, deposited articles etc. The project has started in the middle of 2004 and is funded by DFG and RFBR. The general need to establish such a project is the same as for the DML.

RusDML is planned in accordance with the basic requirements for the DML: The archive should be open and accessible world-wide. Distributed copies should guarantee the safety of the data and facilitate the access from different parts of the user community. DML should be established as global network providing access to interlinked digital publications in mathematics. As a result of the DIEPER project a first sample issue for RusDML even will be available in advance to the project itself. The contents of the most traditional Russian journal in mathematics, *Mat. Sbornik*, had been scanned by the DIEPER partner in Helsinki, and after some additional work on the access data this journal will be available as a RusDML prototype due to kind agreements with the Helsinki University Library for using their files and the Russian Academy of Science (RAS), Moscow Branch, to make the digitised articles freely accessible through the web.

As a key issue the Russian-German cooperation between several partners in both countries will be the organizational base of the project. Scientifically this is a consequence of the traditional good cooperation between Russian and German mathematicians for three centuries. As everybody knows, this cooperation survived some political catastrophes. But even now, when we have a period where Russian mathematicians partially try to publish in other languages, there is a comparatively high interest in Russian publications in Germany and other European countries. It is no question that for Russian mathematicians the digital offer to be installed with RusDML will be a highly desirable improvement of their literature supply. But the two libraries involved on the German side still have the image to be reliable and comprehensive reference sites for this. Providing the content of RusDML will make them unique sites for users who are not likely to go to a provider in Russia. As a consequence, a bilingual access structure with enhanced facilities for those with weak Russian reading capabilities will be one of the most important requirements for RusDML. The conceptual and technical background for this also will be a pioneering tool for digital offers of other Russian publications.

The main Russian partner is the Russian National Public library for Science and Technology (GPNTB). On the level of scientific advice for the project the advice of the Mathematics Division of RAS will be important. Interests of other Russian libraries like those from Kazan, or MGU, or RAS will be respected in bilateral agreements. RusDML will establish GPNTB as the centre of excellence for digital offers of Russian mathematics. But in spite of this there is no aim to interfere with the interests of other mathematical libraries in Russia. Hence their aims and mission will be respected and taken into account, when delivery of documents, linking of offers and mirroring services should be taken into account. GPNTB is the major library in Science and Technology for the Russian Federation. Its collection comprises 8 million items of national and foreign publications. The library provides comprehensive access to Russian collections in its role as a State Depository and recipient of obligatory free of charge copies (mandatory copies) of all publications in their domain. All journals selected for RusDML are available at the collection of GPNTB, starting just from the first issue of the journal in its first publication year to the current production. Moreover, the library is experienced in library automation and information technologies. The Russian Academy of Sciences organizes born electronic offers of their journals and provides them freely for Russian users through their IZIR system. Like with *Mat. Sbornik* they will consider the digitisation of their journals

---

as an added value, and customize everything in a convenient way for their users. In this sense RAS clearly supports the RusDML initiative and this will be an extremely helpful assistance for the negotiations with other Russian publishers and editors for getting the license to digitise their materials. Concerning the scientific exploitation of the RusDML RAS may play a leading role to explain the many advantages of the enhanced digital offers.

---

## 9. The German Part in RusDML

---

The German partners are the State and University library in Goettingen (SUB), the Technical Information Library in Hannover (TIB), and the Technical University Berlin (TUB) representing the contributions from Zentralblatt MATH through its editor-in-chief given by the second author. The different roles of these partners will be explained below. But as an essential common facility it had been agreed, that all three libraries, GPNTB, SUB and TIB, have the option and almost the obligation to install a full copy of RusDML. All project participants are well prepared for the collaboration, because they have pretty good collections in mathematics and they have long experience with the handling and administration of electronic offers.

Zentralblatt MATH has a very special role in this cooperation. Its main duty is to provide a comprehensive reference data base in mathematics. It provides bibliographical data, indexing information and reviews or abstracts in English. Hence the core metadata for RusDML will be available there, because all journals in RusDML are evaluated by Zentralblatt, and employing the linking facilities from the database to full text offers, it can be used as a simple access tool to the holdings of RusDML. The idea is to integrate the reviews as a special addition into the metadata. Hence also users with low reading capability in Russian can decide, if they really want to go into the details of an article or not.

As mentioned above, there are some basic requirements for the project. Most importantly the digital archive should be easily accessible world-wide. All three participating libraries should spend combined efforts to take care of the long-term preservation and readability of the digital collections. Later upgrades of the offers can be imagined leading to more convenient access to electronic archive and to improved search facilities. For example, one important item is the linking from the references to their web offer. To achieve these goals all partners supposed to share their efforts and results and as a consequence they should serve as mutual mirror sites for the complete archive of RusDML. The participating libraries will support international standards for RusDML as recommended by the DML project and others. This keeps the project open for cooperation with other initiatives to be capable for further expansion to cover the Russian publications more completely. There may be additional digital collections provided by other institutions, which are not in the core list of documents recommended for RusDML, for example.

With respect to content several stages of the project have to be considered. On the first stage RusDML will start with processing journals from a core list of about 120 titles, which are covered by bibliographical databases of Zentralblatt MATH and the Jahrbuch database. But in addition to this a big variety of Russian publications in mathematics is available. Hence the core list should be corrected and extended. For instance, in the collection of GPNTB some quite interesting mathematical journals can be found which are not listed as journals in the Zentralblatt database. For this a list of additions has been developed, possibly containing as a journal, which has been classified as a series of collections of articles by Zentralblatt. A registry of Russian publications in mathematics will be developed and extended during the project period.

Coming to the total amount of work in the first stage the following figures have to be considered. Starting with the approximately 120 Russian journals of the first list, which have been processed by Zentralblatt and the Jahrbuch, joint estimates by GPNTB and SUB came to a figure of about 2 million pages. Using the existing digitisation infrastructure at both sides it is agreed that the handling of the structural metadata, which are necessary for controlling the page numbers and the scanning are shared at equal parts between GPNTB and SUB. This delegates about 1 million pages to GPNTB. This will be done on the basis of uniform formats and common protocols with respect to technical issues.

Another nontrivial problem for the content part is the work on licences and permissions. To convince publishers and editors to make their printed publications available on the web needs a lot of promotional work. Some will agree immediately, others will have a lot of reservations and prefer to wait what happens with the first set of publications in RusDML. Thinking about contacting every author will make the work tedious and increase the efforts considerably. The copyright discussion is an open matter in this area, but librarians will have intermediate solutions to survive with an offer in the net in a more or less legal way.

## 10. Added Value of RusDML

Having established a digital version of a journal for its complete publication period will be a first step only. According to the work described above there will be a service related access and navigation structure enabling readers to browse the offer and to read (and print) the text of the articles.

The automatic generation of reference links is likely to operate very soon. The references could be distinguished from the other text in the image. Applying OCR to that part will be able to get structured text information from each reference. Flexible look up systems will contact reference databases like Zentralblatt MATH and arrange a search for exporting the identifier of the reference in order to add this to the metadata of the document under consideration. This can be used to arrange the links from that document to the complete text of the reference, if a digital version is available. Clearly, an adjustment of these tools will be required which can handle Cyrillic characters.

Going beyond these formal procedures dealing with the scanned image itself, the editors and mathematicians interested in the corresponding journal will have the possibility to enrich the information associated with that journal. The static sequence of images and metadata produced for a journal will be capable to store comments, historical remarks or any kind of addition, which seems to be of interest in relation to the scientific merits of the corresponding article. Hence there will be a dynamic aspect in the management of a journal turning even so retro-digitised part into a living archive. This input will be subject to the initiative and the control of the editors. It will provide an improved view of the location of the journal in information space and of its role in the development of mathematics.

This will be an added value for the journal, and it only can be obtained in a convenient way, after having the journal digitised, and equally important, after having provided a structure where useful additional information could be handled in a searchable way.

## Bibliography

- [Andrea Asperti; Bernd Wegner] MOWGLI – A new approach for the content description in digital documents. Ninth International Conference "Crimea 2002" Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 8-16, 2002, Volume 1: 215-219.
- [Bernd Wegner, ELibM] ELibM in EMIS – A Model for Distributed Low-Cost Electronic Publishing. Eight International Conference Crimea 2001O Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 9-17, 2001, Volume 1: 317-320.
- [Bernd Wegner, ERAM] ERAM – Digitalisation of Classical Mathematical Publications. Seventh International Conference Crimea 2000O Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 3-11, 2000, Volume 1: 268-272.
- [Galina A. Evstigneeva, Andrei I. Zemskov] RusDML – A Russian-German Project for Establishing a Digital Archive of the Russian Mathematical Publications, Lecture Notes in Computer Science, vol. 2730, Springer-Verlag, 2003
- [Gerhard Michler] How to build a prototype for a distributed digital mathematics archive library. Proceedings MKM 2001, Linz, <http://www.emis.de/proceedings/MKM2001/>.
- [Hans Becker, Bernd Wegner] ERAM – Digitization of Classical Mathematical
- [Joachim Heinze] Presentation at the first EMANI workshop in Heidelberg, February 2002 (article to appear in the Proceedings of the EIC-Satellite Conference to the ICM 2002, Tsinghua University, Beijing)
- [John Ewing] Twenty Centuries of Mathematics: Digitizing and disseminating the past mathematical literature. [http://www.ams.org/ewing/Twenty\\_centuries.pdf](http://www.ams.org/ewing/Twenty_centuries.pdf)
- [Pierre Berard] Presentation at the San Diego DML-meeting, Joint Mathematics Meeting, January 2002 (see also <http://www-mathdoc.ujf-grenoble.fr/NUMDAM/>).
- [Victor Albis] Conservacion del patrimonio matematico colombiano. <http://www.accefyn.org.co/historia-matematica/histmatcol.htm>;  
<http://www.accefyn.org.co/proyecto/conservacion.htm>; <http://168.176.37.80/matepro.html>  
 Publications, Proc. ECDL 2000, Lecture Notes in Computer Science 1923, 424-427 (2000).

## Author Information

**Bernd Wegner**, Prof. Dr. – Fakultät II, Institut fuer Mathematik, TU Berlin, Sekr. MA 8-1, Strasse des 17. Juni 135; D – 10623 Berlin, Germany; E-mail: [wegner@math.tu-berlin.de](mailto:wegner@math.tu-berlin.de)



---

## MANUSCRIPT DIGITIZATION AND ELECTRONIC PROCESSING OF MANUSCRIPTS IN THE CZECH NATIONAL LIBRARY

Zdeněk Uhlíř

*Abstract:* The paper informs about the history of manuscript digitization in the National Library of the Czech Republic as well as about other issues concerning processing of manuscripts. The main consequence of the massive digitization and record and/or full text processing is a paradigm shift leading to the digital history.

*Keywords:* manuscript digitization, processing of manuscripts, digital history, paradigm shift.

---

### Introduction

---

More or less systematic digitization of manuscripts and other historical materials started only ten or fifteen years ago. In the case of the National Library of the Czech Republic [3] it was in 1992 when the cooperation started within the UNESCO programme Memory of the World. The National Library of the Czech Republic has accomplished a great progress since these sheepish beginnings so that now its digitization team placed itself among the most advanced teams worldwide. The acquired experience during the twelve years shows that the large and massive digitization of manuscripts and other historical materials means a big challenge not only for the computer science and library and information science but for the history in large sense as well because it leads to the paradigm shift in general. This paper concerns some of these issues.

---

### From Manuscript Digitization to the Digital History

---

As I have already said, the start of digitization in the National Library of the Czech Republic is dated in 1992 when cooperation was linked-up with the UNESCO programme Memory of the World. In 1993 the Czech National Library published a pilot CD-ROM [7] for this programme that was created – in close collaboration with the Albertina ICOMA Ltd. [1] – as an example for similar activities in this framework. The pre-history of the digitization activities began at this point that consisted in learning from more developed and advanced teams and institutions as well as in gathering own experience. The results of this development released in 1995 were two pilot CD-ROMs of the newly-created Czech National Library's programme called *Memoriae mundi series Bohemica*, i.e. *Antiphonarium Sedlecense* [2] and *Chronicon Concilii Constantiensis* [4]. It was really the crucial experience leading to recognition that the biggest error concerning digital processing is to put together the data and the software. Thus, the main enlightenment following the first independent result was that the data and the software must be strictly and consequently divided. It is in a clear divergence from the old relational database tradition.

The effect of this recognition was creating the SGML based DOBM standard (Digitization of Old Books, Manuscripts, and Other Materials) that is a document type definition enabling making complex digital documents, i.e. compound documents of bibliographic and technical description as well as images-copies of the original documents. It enabled starting of a large and massive digitization for the programme *Memoriae mundi series Bohemica*. It is very important that following such an achievement on the one hand the programme *Memoriae mundi series Bohemica* became a national programme in 1998 and on the other the DOBM standard was accepted in 1999 as a UNESCO recommendation for its programme Memory of the World. The pre-history finished and the digitization in the National Library of the Czech Republic stepped in its history.

History of digitization in the National Library of the Czech Republic consists in efforts at its webatisation because the net environment is the biggest challenge of the manuscripts work during last several decades. The first step in the effort was an attempt to make available via Internet a special text catalogue [9] that was inspired by the *In principio* in 1998, i.e. text catalogue created for a long time by the French Institute de Recherche et d'Histoire des Textes and the American Hill Monastic Manuscript Library. There were selected ten fields (library, shelf mark, leaves, rubric, first incipit, second incipit, explicit, structural overview, bibliography and note-commentary) that should be able to characterise unambiguously each text item within the manuscript. Perhaps it was a good solution from the point of view of the content but problems came round the technical solution: the traditional

relational database that was chosen however worked very slowly so that it was unacceptable for the final user. Through this negative experience it was discovered through this negative proof that another solution must be chosen. Such a solution consists in the use of markup language which enables representing not only a formal but also a content structure. In other words it means that all branches of manuscript work should be treated together or jointly. This was very important ascertainment.

Thus, in 1999 the National Library of the Czech Republic became one of partners of the European MASTER project (Manuscript Access through Standard for Electronic Records) [13]. The other partners were Humanities Computing Unit of the Oxford University, Center for Technology and the Arts of the De Montfort University Leicester, Institute de Recherche et d'Histoire des Textes in Paris, Dutch Royal Library at The Hague and Arnamagnaeen Institute of the Copenhagen University. Representatives of various, perhaps quite different schools of manuscript work collaborated together in order to create an electronic record for manuscript descriptions. [12] The first idea was to prepare the records in SGML but the writing in XML appeared better at last. The new document type definition was ready by the end of the MASTER project in 2001 and it was widely disseminated through Europe. The importance of the MASTER DTD consists in two things: firstly, it enables preparing short and in-depth record using one document type definition only, and secondly, an extension of the descriptive record facing the complex digital document is possible. The National Library of the Czech Republic started using it instead of the older DOBM standard because it is created exactly according to its long-term needs. The MASTER+ extension enabling connecting the manuscript record with the interrelated digital documents was done in 2002. History was brought to a close and the present started off.

Very important practical and organisational consequences followed after the creation of MASTER+ extension. As it makes possible creating whole complex digital documents, it enables building not only simple web presentations but a true digital library too. And digital library is not any ordinary resource, it is like a gate into the emerging virtual environment. Such a virtual environment must be understood in two simultaneous ways. Firstly, it is a net of newly originating institutions that are different and distinct from the traditional „stone“ institutions as we know them from the modern era. Such virtual institutions are in all probability consortia of traditional institutions that have some new goals, i.e. not only to preserve, conserve and lend the collection items but more likely to present the collection items in over-collection way and to re-present them from various points of view and in different sights. Thus, whichever virtual institution has different tasks in comparison to traditional „stone“ institutions. Secondly, it is a fluid compound representation of transient documents so that resource, not document appears at the first horizon. That means, not individual, but aggregate, collective phenomena are fundamental in such an environment. Very hard consequences follow within the information, communication, and knowledge sphere. Step by step a modern idea of objectivity is replaced by a new concept of virtuality. We are in the model of this great process and we do not know now what it will bring in the end but it is the biggest challenge of our present and near future.

Following these substantial ideas, the National Library of the Czech Republic initiated a decision to put together several of the most active institutions that take part in the Czech national project *Memoriae mundi series Bohemica*. Seven partners (apart from the Czech National Library also National Museum in Prague, Moravian Land Library in Brno, Research Library in Olomouc, Castle Library in Kynžvart, Museum of East Bohemia in Hradec Králové, Praemonstratensian Canony at Strahov) assembled and founded the Memoria project [14] in the end of 2003 that consists in collaboration in developing the virtual research environment for the work with historical holdings. Thus, the Memoria project is actually a consortium of institutions endeavouring to take step from the traditional information, communication, and knowledge environment into the virtual one, i.e. it is the genuine virtual institution. The Memoria consortium undertakes the Manuscriptorium database, [11] founded few months before in 2003 by the National Library of the Czech Republic and maintained by the Albertina icome Ltd. Manuscriptorium database is from its very beginnings oriented to the cooperation and integration with a wider circle of foreign partners. University Library Bratislava in Slovakia became the first one. Other partners that signed agreement with the National Library of the Czech Republic as coordinators of the Manuscriptorium database on the manuscript cataloguing are the University Library Wrocław in Poland and the National and University Library Zagreb in Croatia. Currently testing of a more sophisticated integration is running. It consists in joining of the results of the Austrian project *Monasterium* [15] and the German one *Codices electronici ecclesiae Coloniensis*. [10] The *Monasterium* database this way represents an important attempt to integrate the cultural heritage at the supranational, transnational level for Central Europe.

---

Thus the main idea of the Manuscriptorium database is cooperation and integration. There are several aspects of this idea. Firstly, it concerns various organisational levels: the Czech National Library programme (Memoriae mundi series Bohemica); the Czech national programme (Libraries' Public Information Services, branch 6); open group of the „willing“ partners within the Czech national programme (Memoria); and finally open group of the foreign partners (now Slovakia, Poland, Croatia, Austria, Germany – and the others welcome: at the moment we are negotiating with the Library of the Lithuanian Academy of Sciences in Vilnius). Secondly, it concerns various material types (manuscript books, incunabula, early printed books, maps, graphics, charters, etc.), i.e. it is the idea of interdisciplinarity and transdisciplinarity that is the most important challenge of contemporary work with historical materials. Thirdly, it concerns integration of various document types (catalogue records, digital replicas/copies – images, sounds alike), fulltexts of primary, i.e. original, and secondary, i.e. interpretative, documents, eventually multimodal documents that are substantially compound and transient ones. And fourthly, it concerns integration of cultural heritage at the transnational Central European, eventually even European level. Thus, the main and proper goal of the Manuscriptorium database is to create a gate for the manuscript and other historical sources studies in a global dimension.

Now, it is very important to know how to do it in particular. The first step is to use the MASTER and MASTER+ records. Records created according to this double standard enable a goal-directed choice between the short records and on the other hand the in-depth records and subsequently to create a reasonable time-management. The MASTER records also enable the choice among various kinds of manuscript description according to the various purposes that the records are procured for and subsequently to aspire to interdisciplinarity and/or transdisciplinarity. Another important characteristic of the MASTER+ records is that they enable connecting the descriptive record with the appropriate interrelated digital documents and in this way making complex documents that are compound as well as transient, i.e. to build the digital library as a basis for the global virtual research environment. Last, but not least the use of MASTER and MASTER+ records is the necessary condition for interoperability because its' consequence is the clear and apparent divide between the data and the software. Although to learn to create and to use the MASTER and/or MASTER+ records is a hard work the results of it are well-arranged and user-friendly.

The second step is to use standards for creating and processing images. These standards guarantee that the images will be of an excellent duality and subsequently that they can be archived and used again. The use of such standards enables making them accessible for browsing and/or searching according to the user's purpose/s and subsequently building and providing an indirect service that is the keystone of the electronic, i.e. net and virtual environment. On the other hand, it enables building and providing a direct one which consists in digital reproduction delivery services. Such a possibility of combination of direct and indirect services is the greatest advantage of the electronic environment in comparison to the traditional environment of the printed book. In a further perspective using standards for creating and processing images leads to the possibility of building and providing various levels of indirect services according to the various quality levels that are made on the basis of various types of conversion, compression and so on and so forth. Of course, in this case there is some kind of authentication and licence management needed. It is a big challenge because contemporary copyright law as well as ideas about the intellectual property do not have at any case friendly inclination neither to the electronic resources nor to the electronic environment.

The third step is to use a purpose targeted adaptation of the TEI standard. [17] It can be generated automatically using the Pizza Chef. [16] The TEI standard and its derivations enable creating various kinds of full texts concerning form of markup as well as content, i.e. it is able to process editions of the original documents on one hand and secondary documents (documents about these editions and/or documents about other documents, facts, events, persons, artifacts, etc.) on the other. Thus, the TEI standard follows the compatibility of all such documents and it enables subsequently very easy and comfortable archiving. The TEI document type definition makes possible a choice among various "markup ideologies" because of its flexible content based markup, i.e. among various even different purposes of the created document. That means not only the fulfillment of the interoperability requirements but also the possibility to do simple transformations and subsequently to make documents easily accessible and the possibility of a simple way to e-publishing. The consequence of compatibility and interoperability at this level is far-reaching. It means not only an implicit interdisciplinary and/or transdisciplinarity but also the possibility to evolve its' explicitly. Upon the basis of only one archive database there are many presentation databases that can be created, maintained, and provided. The difference between

individual presentation databases consists in the possibility to eke out the markup of document existing within the archive database with other specific markup according to the same or another TEI standard adaptation. It is a big advantage in comparison to traditional printed representations as well as a challenge to think in another ways than usual.

The fourth step is to use connectivity standards as Z39.50, OAI PMH alike. It enables integrating mutually various resources of the same document type, i.e. the primary documents (original historical documents and their various representations), the secondary documents (documents about primary documents and other interpretative ones), and the tertiary documents (catalogues and/or bibliographic records, documentation items alike) that use some kind of metadata (which stands to a reason nowadays). The consequences are the resource's interrelations and the completion of information. It enables creating indexes at the over-resource level, too. Such indexes facilitate heuristics within completed as well as individual resources. It is a welcome and important contribution for solving problems that so called second information crisis brought along. The connectivity protocols build an initial level of the virtual research environment because of easy orientation and navigation. Of course, even though it is very inchoate it is a step in the right direction.

The fifth step consists in the simultaneous, flexible use of various approaches. The most important thing is not to seek solitary ways but to use standardized and regularized tools of the net environment as Internet browsers, markup language editors and processors, parsers, validators, coding and format convertors alike. On the other hand together with the use of these regularized and standardized tools is quite a clear need to use and/or to create individual specific *ad hoc* applications as computational linguistics tools and systems etc. The multitasking facility of personal computer is another possibility how to do the work in the electronic environment more reasonably and user friendly because it enables using computer as a desktop in a quite real, not only metaphorical sense. Now, there are many and many such tools and applications at our command so that we can see entirely clearly that we need "new" methodologies instead of the "old" ones. Thus, although at the first sight it looks as if the easiest and simplest step according to all experience is the most difficult one. To seek "new" methodologies means herewith to develop a new paradigm, that is quite different from the old one, so it is a sure way to the uncertainty because it disrupts the previous certitude. Nevertheless if we go through this way step by step, we have learned and got to know that the paradigm shift comes slowly but surely.

The recognition follows that the issue of the virtual research environment for the work with historical holdings and cultural heritage is a question of its content, not of technics and/or technology. Thus, although technics and technology is *conditio sine qua non*, it is an insufficient condition. That means, as so as the information-communication technologies are an aid only and not the goal, the main problem concerns the content. The most important problem of the content is that the content presented as a cultural and scientific heritage is still the content of yesterday, not of today and tomorrow. So it must be adapted according to the requirements of the information wave and information society, not according to the ideas of the industrial ones. There is a crucial need to investigate which is the difference – we might say the *specific difference* – between these two fundamental conceptualizations, before it will be possible to develop any feasible and reasonable new ideas.

Of course, this problem is very large and difficult to be described or characterized in general. [18] So it must be narrowed for the domain of historical materials, i.e. cultural and scientific heritage only. The traditional conceptualization of cultural and scientific heritage within the industrial, i.e. modern society is based upon the idea of objectivity [8] of the world and of things within the world and subsequently upon the idea of the artifact. The idea of objectivity has been gradually replaced by the idea of intersubjectivity since only several decades ago. Artifact is not a simple and indifferent cultural object; it is a consciously and willfully created work. So the fundamental idea of cultural and scientific heritage is the idea of the work. The conceptualization of the work follows two main points of view, the historical and the philological. From the point of view of history *sensu largo* it is based upon the idea of external features in the sense of the *historische Hilfswissenschaften*. From the point of view of philology – and eventually also art history – it is based upon the idea of individuality and artificiality. The work according to the ideas of the industrial society is a real material thing, an external material object, not an ideal conceptual subject, an internal mental object. The work in this conceptualization simply is what is left, not it that it must be understood. The work in this sense is conceptualized as the *capta*, i.e. recorded and preserved data; it is not information in any meaningful sense. The understanding of cultural and scientific heritage within the industrial society is a consequence of that, a life in the industrial society is based on the operations with things,

*atoms*. On the other hand, the arising information society is based on the operations with the signs and/or symbols, *bytes*. The difference between these two conceptualizations is a paradigmatic difference. Thus, because of different paradigms, i.e. different discourses there are different methodologies that construct different "facts". For the industrial and information wave/society the "facts" are not the same. That means, the "fact of yesterday" is not the "fact of today" and far less the "fact of tomorrow".

This is a real and difficult problem that must be solved if we want to go further. There are two natural ways how to solve this problem and they are given to us simultaneously. The first of them follows the recognition that objects, i.e. artifacts. Works are no external real objects but internal ideal objects; they are objects of human mind. They are some representations of external objects only. As they are representations they are herewith interpretations, explications, explanations, imaginations, fantasies, etc. Thus, as for cultural and scientific heritage, the paradigm shift doesn't concern as much preservation of the real objects as but rather preservation of such mental objects. That means that there is no need to preserve the all what was achieved but all what is apt for preservation. On the other hand, there is a question what it means, when we say "to be apt for preservation". If we agreed that cultural and scientific heritage is not solely the realm of things, i.e. not of the external objects, but more likely ideas, i.e. the internal subjects, then the preservation concerns ideas, not merely things. We must preserve especially ideas, not things, not the discourse concerning things and operating with ideas. As so as this discourse is the "discourse of yesterday" that operates with the "facts for yesterday". Thus, the goal of the work with historical materials, with cultural and scientific heritage is to create "facts of today" with the "regard for tomorrow".

The consequence of this recognition is that the path to the virtual research environment is not just a simple conversion from the traditional printed environment to the electronic one; it is not a transformation of printed representations to the electronic ones (and far less the mere retro-conversion). The first main issue of the path to the electronic virtual environment is to find new forms of representations so that they diverge from the printed ones. The complex digital document (that is the essence of the virtual environment that consists in the evidence record connected with the interrelated documents) is no simple accumulation of individual partial documents but it is a structured net of documents that are transitional, i.e. they generate the integrated transitional compound document. [21] This could mean that such a complex document doesn't always need to have the same constituents and that the constituents can differ in various request situations. If so, virtuality is given at least potentially. The question is how to bring potential virtuality into the actual one. Thus, the other issue of the path to the electronic virtual environment concerns methodologies and facts created according to the new methodologies, i.e. it is the issue of paradigm shift. The question of the methodology as well as paradigm is the crucial question now.

The first substantial step during the way to the virtual research environment for the work with historical holdings and/or cultural and scientific heritage must be a change of understanding of cataloguing, bibliography, documentation, etc. The previous understanding of these activities and subsequent products is based on the typical situation of printed publications because it comes out from the external features and the idea of the text as work. It does not correspond with the typical situation of manuscripts and dominant manuscript environment. It does not correspond with the typical situation of electronic or digital documents. [20] Now such an understanding of cataloguing, bibliography, documentation, etc. is crucial which is oriented to the internal features in the sense of the *historische Hilfswissenschaften* [19] and to the idea of the text as floating continuum, [3] not as the work. Requirements for bibliographic and other similar records deviated from the manifestations (publications, editions) and items (copies, holdings of publications) to expressions and works in a quite virtual sense. [5] Subsequently there is a real possibility that the same (i.e. identical) item as well as manifestation (text, document) can be part or constituent of various expressions and works so that we must accept a new type in these scales, the work expression that is not simple accumulation of the expression and the work. It is a fully new domain of knowledge that must be seriously searched.

---

## Conclusion

---

Thus, the digitization team of the National Library of the Czech Republic asks what to do in the near future. There are three fundamental tasks: first of all, transferring all the catalogues, bibliographies, documentations, factual and material studies as well as historical text editions into the virtual, i.e. electronic, digital environment; second of all, creating many information communication technologies tools for the mass processing of historical documents

---

and holdings; and third of all, starting creating of the multimodal resources for presentation and re-presentation of the integrated cultural and scientific heritage. It will take some years – and then we will see furthermore.

---

### Bibliography

---

- [1] Albertina icome, see at the URL <http://www.aipberouon.cz>.
- [2] Antifonář Sedlecký. Antiphony of Sedles. Antiphonarium Sedlecense, MS XIII A 6, ed. Zdeňka Hledíková – Hana Hlaváčková – David Eben. CD-ROM, Praha, Národní knihovna & Albertina icome Praha, 1995.
- [3] Bryant, John: The Fluid Text, Ann Arbor, 2002.
- [4] Chronicon Concilii Constantiensis. Malá Riechentalova kronika, MS VII A 18, ed. Zdeněk Uhlíř. CD-ROM, Praha, Národní knihovna & Albertina icome Praha, 1995.
- [5] Functional Requirements for Bibliographic Records: Final Report, München, K.G.Saur, 1998, available at the URL <http://www.ifla.org/VII/s13/frbr/frbr.pdf>.
- [6] National Library of the Czech Republic, see at the URL <http://www.nkp.cz>.
- [7] Paměť světa. Mémoire du Monde. Memory of the World. CD-ROM, Praha, Národní knihovna v Praze & Albertina icome Ltd., 1992.
- [8] Popper, Karl Raimund: Ausgangspunkte. Meine intellektuelle Entwicklung, translated Friedrich Griese, München, Piper Verlag, 2004; Popper, Karl Raimund: The Open Society and Its Enemies, London, Routledge & Kegan Paul, 1962; Popper, Karl Raimund: The Poverty of Historicism, London, Routledge & Kegan Paul – Boston (Mass.), The Beacon Press, 1957; Popper, Karl Raimund: The Logic of Scientific Discovery, London, Hutchinson & Co. – New York, Basic Books Inc., 1959.
- [9] Text Catalogue, see introduction available at the URL <http://digit.nkp.cz>.
- [10] The Codices electronici ecclesiae Coloniensis project, available at the URL <http://www.ceec.uni-koeln.de/>, eventually <http://www.ceec2.uni-koeln.de/>.
- [11] The Manuscriptorium database, available at the URL <http://www.manuscriptorium.com>.
- [12] The MASTER document type definition, available at the URL <http://www.tei-c.org.uk/Master/Reference/DTD/>, eventually <http://www.tei-c.org.uk/Activities/MS/FASC-ms.pdf>.
- [13] The MASTER project, available at the URL <http://www.cta.dmu.ac.uk/projects/master>.
- [14] The Memoria project, available at the URL <http://www.memoria.cz>.
- [15] The Monasterium project, available at the URL <http://www.monasterium.net>.
- [16] The Pizza Chef tool, available at the URL <http://www.tei-c.org/pizza.html>.
- [17] The TEI document type definition, available at the URL <http://www.tei-c.org/Guidelines2/index.html>.
- [18] Toffler, Alvin: The Thire Wave, London, Pan Books, 1981; Negroponte, Nicholas: Digitální svět. Being Digital, translated Petr Koubský, Praha, Management Press – Softwarové noviny, 2001; Flusser, Vilém: Do universa technických obrazů, translated Jiří Fiala, Praha, OSVU, 2001; Flusser, Vilém: Kommunikologie, ed. Stefan Bollmann – Edith Flusser, Frankfurt am Main, Fischer Taschenbuch Verlag, 2000.
- [19] Uhlíř, Zdeněk: Teorie a metodologie elektronicko-digitálního zpracování rukopisů a hybridní knihovna, Praha, Národní knihovna České republiky, 2002.
- [20] Uhlíř, Zdeněk: Terminologie a pojmy v čase paradigmatických změn, Národní knihovna: knihovnická revue, 14, 2003 (4), p. 236-244.
- [21] Williams, Robert F.: What's New: Transient Compound Documents Establish Irrevocable Records, available at the URL [http://www.cohasset.com/main/library/coh\\_articles/whatnew\\_body\\_transient.htm](http://www.cohasset.com/main/library/coh_articles/whatnew_body_transient.htm).

---

### Author Information

---

**Zdeněk Uhlíř** – National Library of the Czech Republic, Department of Manuscripts and Early Printed Books, Klementinum 190, 11000 Prague, Czechia; e-mail: [Zdenek.Uhlir@nkp.cz](mailto:Zdenek.Uhlir@nkp.cz)

---

## INTEGRATED AND PERSONALIZED DIGITAL INFORMATION SERVICES

Yaşar Tonta

*Abstract:* Digital information services are gradually becoming integrated with other systems and services such as library automation systems, student information services, and electronic learning systems. Users demand seamless access to a multitude of digital information services without leaving their desktop computers. They prefer using systems that recognize them when they log on, acknowledge their rights and privileges, and thus provide personalized information services. This paper summarizes the recent developments concerning integrated and personalized digital information services. It first emphasizes the role of the Internet in providing information services and then goes on to discuss the integration and personalization issues by emphasizing their importance for digital information services.

*Keywords:* Digital information services, Personalization, Integrated information services, Personalized information services.

---

### Digital Information Services

---

A large number of digital information services are offered to users by library and information centers today. Table of contents services, full-text access to electronic journals, current awareness services, electronic document delivery, and virtual reference services are among them. The Internet plays a paramount role in providing these services. While Internet removes both temporal and spatial barriers and makes it possible to provide information services on a 24/7 basis, it also provides "instant gratification" to users. In other words, users can get what they want instantly, from anywhere, at the best value for their money [Reich, 2002, p. 15].

Information technology (IT) in general, and the Internet in particular, has had a significant impact on information management in that it enabled library and information centers to shift their services from centralized information services to the networked and distributed ones. Consequently, economic models of information provision have also shifted from models based on centralization to that of personalization. The ability to make an information source simultaneously available to multiple users through networks removed the "one source – one user" limitation of the print world. Information centers are no longer bound with their own information resources to provide services as they can easily get access to remote information sources "just in time". While the ownership of information sources in the print world dictates the use of centralized information management models, simultaneous access to the same source by multiple remote users through networks paves the way for more cooperative/consortial information management practices.

The costs of information sources have increased tremendously within the last 20 years. The material acquisition budgets of the members of the Association of Research Libraries (ARL) in the United States can no longer keep up with the accelerated price hikes of information sources, especially scientific journals. Libraries devote about 20% of their material acquisition budgets to electronic information sources. The total expenditures of ARL libraries for electronic serials have increased more than ten-fold between 1995 and 2002 [Kyrillidou, 2003]. Cooperative practices tend to ease the burden of library and information centers and enable them to spend more money to license electronic information sources. At the same time, such consortial initiatives increase the responsibilities of collection managers as they have to prepare separate policies of licensing, processing, maintenance, storage and use for electronic information sources. Moreover, consortial agreements make each library and information center "interdependent" on other such centers, library consortia, information producers/providers and aggregators.

Traditional information management practices are usually based on centralization of resources and services. Access to centralized information sources and services is provided through intermediaries such as reference librarians or information specialists. Intermediation requires centralization. It is expensive; it usually means long lines queuing for centralized services; and it does not serve the remote users. The use of IT and networks on the other hand makes information management less centralized and more distributed. Remote use of resources and services is often carried out without intermediation. Disintermediation decreases in-house use of library materials,

reference and circulation transactions while it increases interlibrary borrowing transactions and electronic document delivery requests. For instance, the number of reference transactions in ARL libraries was declined almost one-third while the inhouse use of libraries was almost halved since 1991. Interlibrary borrowing, on the other hand, has doubled during the last decade [Kyrillidou, 2003]

---

### **Integration of Digital Information Services**

---

Nowadays, digital information services are often integrated with other computer-based systems. For example, digital information services are usually blended with library automation systems, student services, financial services, research and grants data management systems in universities. More often than not, digital information services have links with off-campus electronic learning, electronic government and electronic commerce systems. Interoperability makes it possible to integrate digital information services with other on- and off-campus information systems.

In the recent past, a library has usually provided access, by means of different user interfaces, to its online catalog, bibliographic and full-text databases, table of contents services, electronic preprint archives, citation indexes and the resources available through the Web. Users had to invest time and energy to learn how to use each interface with differing modes of interaction and to study each database with differing record structures, metadata schemes, etc. in order to carry out successful searches. The recent availability of commercial, off-the-shelf software packages such as SFX enabled libraries to integrate different digital information services and offer access to them through a single interface. Such systems link different databases that are available in the library and create more complete records that would likely satisfy users' search requests. For instance, certain parts of a bibliographic record representing an item may come from a citation database, library's online catalog, serial holdings, subject gateways or reference databases such as Ulrich's or PubMed. A library may have access to the same source in different formats (e.g., printed, CD-ROM, on-line). Each format may have some advantages and disadvantages. While the on-line copy may be the most convenient from the users' point of view, it may not be the most cost-effective from the library's point of view. If a source is available in more than one formats in the library, such software packages can also handle what is called the "appropriate copy" problem based on library's specifications. Furthermore, link software packages are usually integrated with the library automation systems (e.g., Millennium of the Innovative Interfaces, Inc.) that take care of routine maintenance tasks such as acquisition, cataloging, and circulation.

---

### **Personalization of Digital Information Services**

---

Personalization is defined as ". . . selecting and filtering information objects or products for an individual by using information about the individual" [Koch, Möslin and Schubert, 2002]. It became cheaper to produce personalized goods and services using advanced IT. Toffler pointed out that ". . . as technology becomes more sophisticated, the costs of introducing variations declines [Toffler, 1970, p. 236]. In fact, mass customization and personalization is an indication of a rich and complex society (Information Society) whereas mass production and mass distribution has been one of the identifying characteristics of the Industrial Society. The Industrial Society is based on what Mitchell M. Tsang called "make, store, sell" approach while the Information Society promotes the "sell, make, deliver" approach. In other words, producing faster, cheaper products in large quantities has been the cornerstone of industrial societies. On the other hand, an idea or a product is first sold to a customer and then it is developed according to the customer's specifications and finally delivered to the customer.

Internet users should be familiar with personalized information services such as personal banking services, on-demand publishing and on-demand video services, automatic current awareness services, electronic document delivery services, recommender systems, and personal information agents. The availability of personalized services makes life easier for us as it is usually more convenient to get access to, say, our bank accounts through networks from wherever we happen to be (home, work place, etc.) instead of paying a visit to the bank.

In order to personalize information services, personal information about users (their demographic characteristics as well as their information seeking and use behaviors) needs to be gathered. This can be achieved either implicitly or explicitly. For instance, several search engines collect personal data about users by means of "cookies" or by using click-stream analysis techniques. Or, users may take a more active role in personalization and define their interests by voluntarily filling in web forms. Once the user profiles are created by using both



implicit and explicit methods of data gathering, "pull" and "push" technologies can be used to provide personalized digital information services.

Personalization can be applied in three different levels. Most web users are familiar with the personalization of the display environment of such information services as My CNN, My Yahoo!, My Bank, or My Library. Users themselves can easily specify where each item or icon should appear on the computer screen. The content can also be personalized based on user data or specifications. For example, users can choose the type of news (sports, politics, etc.) they wish to read when they connect to the system or the weather forecast for their geographic region. Personalization of content is also used in library and information centers. The digital collections available through the library can be made accessible to users based on their statuses (e.g., student, academic) or their origin of network connection. For instance, the use of digital collections of a library is often restricted to licensed users only. Remote users connected to the library from computers outside the specified IP domain(s) may not access to some of the digital resources even though they are available. Or, certain electronic reserve collections can only be accessible by those students taking a specific course. In other words, "availability" and "accessibility" are two different things. The digital content can be personalized according to users' characteristics. The digital information services such as "alert" or electronic document delivery services can also be personalized. The personalization of services requires a more sophisticated approach. Not only does the digital content get personalized but also the services need to be tailored according to each and every user's rights and privileges. For example, if a library does not own or has access to a certain information source, a professor may place (up to a predefined threshold) electronic document delivery requests using the facilities provided through the library automation system whereas a student may not be able to do so or may do so only through the intermediation of an information professional.

A complex matrix defining the users' rights and privileges as well as collections' characteristics (e.g., open to campus users only, or open to students taking a specific course) needs to be created beforehand to provide personalized digital content and digital information services. Online bookstores such as Amazon.com keep data about past browsing and buying behaviors of their clients so that they can recommend new items to consider/buy when the user logs on to the system next time. Although library and information centers can easily gather similar data about users' past borrowing or downloading behaviors and use this data to provide more personalized services, they refrain from doing it for various reasons. First, they may not keep personal data longer than a specified period of time according to the law and the personal data may only be reported in aggregates and destroyed later. Second, some users may feel uneasy about being "observed" by the information providers. Third, the library automation systems are usually not configured to relate content and service data with personal data. Lynch [2001] points out that "circulation systems typically break the link between a patron and a book that has been borrowed when that book is returned" and thus libraries lose the opportunity of providing more personalized services. Lynch also emphasizes the fact that it is quite difficult to implement personalization in a distributed information environment as personalization "occurs separately within each system that one interacts with" and "[i]nvestment in personalizing one system (either through explicit action or just long use) are not transferable to another system." However, personalized digital information services should be provided to those who are willing to experiment. There should be nothing wrong with sending a personal message automatically to a user's mobile phone informing her that the article she has requested earlier is now available or sending an electronic copy of the article to her electronic mail address.

Information centers are increasingly designing "portals" to facilitate the users' access to the available resources. A portal can be defined as an information hub or an entry point to digital information sources and services available through the Internet and intranets. It is an application that provides metasearch and support services. A portal can be an entry point to an institution's repository as well as to community repositories and commercial resources. It can provide a number of services ranging from terminology services to rights management, from harvesting data to identity management, and from configuration to presentation. A portal also provides personalized sources based on personal demand or specific roles and organizes the digital content so as to help different users. Portals can be seen as a "one stop shop" because they are designed in such a way that users could find what they want by simply visiting a portal's web site. Yet Dempsey [2003] warns us that, if not designed carefully, portals may also function as "one shop stops," thereby limiting users' choices of access to information. He further adds that a "portal is not a strategy replacing effective use and management of information sources in a networked environment, but, rather, is a part of such a strategy" [Dempsey, 2003].

---

## Conclusion

---

When one gets access to a library web site, it is not unusual to see that the standard content is offered to all users regardless of their rights and privileges. Yet users should be recognized when they log on and the content and services should be personalized accordingly. This can be done using, among others, various “pull” and “push” technologies, smart cards and – in the foreseeable future – biometric features. To do that, library and information centres need to move from “resource-centric” approach to “relationship-centric” approach. To put it differently, library and information centers should take into account not only the relationship between bibliographic records and the information sources that they represent but also the relationship between information sources and the users. As I emphasized elsewhere, “instant gratification” is only possible with the availability of instant access to networked, personalized digital information services. Without this, more demanding users will take their business elsewhere to get instant satisfaction [Tonta, 2003].

Several issues should be tackled in order to be able to offer truly personalized digital information services. Personalization is difficult to implement in distributed environments. It requires networking infrastructure (access to personal, local, regional and wide area networks). It also requires interoperability in that digital information services should be interoperable (i.e., work together) not only with library automation systems, student information systems, financial systems but also with e-banking, e-commerce, e-health, e-government and e-(l)earning systems. Sound policies to tackle security and privacy issues should be implemented along with more sophisticated budgeting, pricing, use and training models. If library and information centers do not succeed in their endeavors of providing personalized digital information services, there is a danger that they might be ignored and neglected by the potential users in the future.

---

## Bibliography

---

- [Dempsey, 2003] L. Dempsey. The recombinant library: portals and people. *Journal of Library Administration* [Online]. Available: [http://www.oclc.org/research/staff/dempsey/dempsey\\_recombinant\\_library.pdf](http://www.oclc.org/research/staff/dempsey/dempsey_recombinant_library.pdf) [17 October 2003].
- [Lagoze, 2000], L. Lagoze. Business unusual: how “event-awareness” may breathe life into the catalog? [Online]. Paper prepared for *Bicentennial Conference on Bibliographic Control for the New Millennium, Library of Congress, November 15-17, 2000*. Available: [http://lcweb.loc.gov/catdir/bibcontrol/lagoze\\_paper.html](http://lcweb.loc.gov/catdir/bibcontrol/lagoze_paper.html) [02 April 2003].
- [Lynch, 2001] C.A. Lynch. Personalization and recommender systems in the larger context: New directions and research questions. [Online]. Keynote paper presented at *Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin, Ireland, June 18-20, 2001*. [Online]. Available: <http://www.ercim.org/publication/ws-proceedings/De1Noe02/CliffordLynchAbstract.pdf> [29 March 2003].
- [Koch, Möslin, and Schubert, 2002] M. Koch, K. Möslin and P. Schubert, Communities and personalization for individual products. In: *Proceedings of the British Academy of Management Annual Conference, London, UK, 9-11 September 2002*. [Online]. Available: [http://dwi.fhbb.ch/eb/publications.nsf/0/25e3487be40f4796c1256be90074c394/\\$FILE/koch\\_158\\_final.pdf](http://dwi.fhbb.ch/eb/publications.nsf/0/25e3487be40f4796c1256be90074c394/$FILE/koch_158_final.pdf) [20 October 2004].
- [Kyrillidou, 2003] M Kyrillidou, E-Metrics: lessons learned from the ARL E-Metrics Project: Challenges and opportunities. Presentation delivered at the 226th American Chemical Society National Meeting, September 8, 2003, New York, NY. [Online]. Available: [http://www.libqual.org/documents/admin/acs\\_kyrillidou.ppt](http://www.libqual.org/documents/admin/acs_kyrillidou.ppt) [20 August 2004].
- [Kyrillidou and Young, 2004]. M. Kyrillidou and M. Young. Research library trends. [Online]. Available: <http://www.arl.org/stats/arlstat/03pub/03intro.html> [19 August 2004].
- [Reich, 2002] R. B. Reich. *The future of success: working and living in the new economy*. Vintage Books, New York, 2002.
- [Toffler, 1970] A. Toffler. *Future shock*. Random House, New York, 1970.
- [Tonta, 2003] Y. Tonta. The personalization of information services. *Information Management Report*, August 2003, pp. 1-6.

---

## Author Information

---

Yaşar Tonta – Department of Information Management, Hacettepe University, 06532 Beytepe, Ankara, Turkey; e-mail: [tonta@hacettepe.edu.tr](mailto:tonta@hacettepe.edu.tr)

---

## DESIGNING A CULTURAL HERITAGE SECTOR BROKER USING SDBC

**Boris Shishkov**

*Abstract: Among the actual cultural-heritage-related problems is the one of effectively managing and globally distributing digitized cultural (and scientific) information. The only feasible way to realize this goal is via the Internet. Hence, a significant issue to be considered is the adequate design of software applications which to realize brokerage tasks within the global space. However, due to the great complexity of this cultural-heritage-related task (compared to other brokerage tasks successfully realized by software systems), the usage of the existing popular modeling instrumentarium seems inadequate. Hence, in this paper, an approach is presented and it is briefly discussed how the approach could be useful for building cultural heritage sector brokers.*

*Keywords: SDBC; Software broker; Cultural heritage*

---

### Introduction

---

As it is well-known, several types of activities are observed, which concern the digitization of cultural and scientific heritage. Among them are:

- the classification of existing cultural heritage materials;
- the recognition and processing of images from such materials;
- the specification and maintenance of the metadata related to digitized materials;
- the management and distribution of the digitized cultural and scientific materials.

In the paper, this last issue will be addressed. In tune with the current technological possibilities and user demands, such management and distribution should be considered in global respect. It is necessary that the digitized cultural and scientific materials are globally available to the public. Next to that, their accessibility must be regulated. This is not easy because some materials are to be accessible freely by anyone, others should be accessible only by authorized users, still others are to be distributed commercially, and so on. Thus, an advanced brokerage functionality needs to be realized in this regard. As far as the 'global information space' is concerned, this job is to be done by a 'software broker'. By 'software broker' it is meant a software application which realizes a brokerage functionality. It is well-known that software brokers exist and are used for a number of purposes, for instance, flight/accommodation reservations, e-Business, Tele-Work, and so on. However, a global management and distribution of digitized cultural (and scientific) data, which is characterized by a number of restrictions (as above mentioned), makes the brokerage task more complex than what is usually observed in brokerage systems. It is, therefore, essential specifying such a (software) brokerage system based on a sound consideration of the original business system, to be supported by it. This leads to a more general actual research problem, namely the alignment between business process modeling and software specification.

Further in the current section, this issue is addressed and afterwards, the software brokerage systems and their relation to the cultural heritage aspect are considered. Software applications are supposed to have the crucial role of an intermediary between the technology (in particular – Information and Communication Technology – ICT) and the (business) processes supported by it. Hence, an essential issue for current business development is the effective application support. Considering the specification of applications that should support business processes, one frequent cause of software project failure is the mismatch between the users' requirements and the actual functionality of the delivered application. Most of the current software design methods are characterized by a lack of proper alignment between the consideration of users' requirements and the specification of a software conceptual model. Actually, we observe two opposite phenomena [Shishkov, 2002]. On one hand, we observe software being developed without prior adequate investigation of the business processes to be supported by it. This means that the business requirements are poorly determined and the software design model is not rooted in a business process model. Therefore, the developed applications would support the business processes inadequately. Hence, although the applications' quality might be high from a

software point of view, the effectiveness of the support they offer to the target business processes would remain low.

On the other hand, although sound business process modeling is often conducted prior to the design of applications, the business process model is only partially used, since it is not straightforwardly transformable into a relevant input for the application design. This does not allow for full employment of the ICT possibilities in solving the particular business problem(s).

Therefore, the two outlined tasks need to be aligned in a better way: the business process modeling and the specification of software applications for the support of the business processes. They both should be considered as one integrated task.

Many researchers have addressed issues related to these problems [Shishkov & Dietz, 2004-2]. Dehnert and Rittgen present a formal representation for describing business processes. This is a promising step and could be especially useful if further related to software design. Olivera, Filho and Lucena have also contributed in this direction, by investigating the design of software on the basis of business requirements analysis. Their suggested approach is a step ahead even though it does not yet offer a straightforward mapping of a business process model into a software design model. Hikita and Matsumoto have studied how the appearance of additional requirements could be reflected in the system's construction, which is also a promising result achieved so far (although not completely solving the problem). Krutchen suggests (based on the existing use case concepts [Jacobson et al, 1992]) a "Business use case" – considered useful in bridging business process modeling and software design [OMG, 2004]. But it is still a question how to consistently identify such use cases. These are only some of the examples of research activities addressing the mentioned problem, having reached no complete and convincing solution. Therefore, it might be concluded that further knowledge is still required in the direction of soundly basing application design on business process modeling.

With regard to the outlined research problem, a promising contemporary approach for application development is the component-based development [Jacobson et al, 1992], founded on the principles of object-orientation (OO). As it is well known, the OO paradigm (characterized by the fundamental concepts of encapsulation, classification, inheritance and polymorphism) is widely considered as a special approach to the construction of models of complex systems, in which a system consists of a large number of objects. According to some researchers, this could be applied not only to software systems but also to business systems [Shishkov & Dietz, 2004-1]. Thus, it seems feasible to expect that software specification and business process modeling could be bridged by basing the specification on software components which are derived from some business components. Such components should fill the gap between the two mentioned tasks. If re-usable components are identified, they could be used many times for designing different applications. Next to that, component-based development seems beneficial for the application design itself. By basing application development on encapsulated, individually definable, re-usable, replaceable, interoperable and testable (software) components, developers could build applications which possess durable configuration and a high degree of flexibility and maintainability. The process of application development would also be improved because building new applications would include using already developed components. This reduces development time and improves reliability. The performance and maintenance of developed applications would be enhanced because changes could occur in the implementation of any component without affecting the entire application. All this makes the component-based application development much more effective than the traditional way of application development.

For all these reasons, in considering the problem of alignment between business process modeling and conceptual software specification, the focus is, in particular, on realizing this on the basis of (re-usable) business components identified from target business processes. By basing the design of applications on such components, it is claimed that the application support to business processes can be improved considerably. A resulting effect from applying a component-based business/software alignment would be improvements in specifying powerful software systems that realize advanced brokerage functionality. Such a functionality would be adequately derived on the basis of the original business requirements. This would make such systems much effective in solving problems in the context of the originally existing business environment.

A software broker, specified in such a way, would be useful for some cultural-heritage-related activities, as mentioned before. They could be significantly facilitated via such an advanced software support. This justifies the current study and inspires the efforts to bring together the cultural heritage domain and latest software design

achievements. Additional motivation brings the fact that no information was found about a cultural heritage project or initiative where the brokerage problem has been adequately solved. In this paper, the SDBC (SDBC stands for Software Derived from Business Components) approach is considered as an adequate software specification tool as far as the discussed brokerage problem is concerned. SDBC is a conceptual modeling approach introduced as a way to improve the quality of the software production process. The approach allows for soundly aligning (in a component-based way) business process modeling and software specification, and could be particularly useful in specifying software brokers. These issues are addressed also based on the consideration of SDBC. Besides this, the particular relevance to the cultural heritage issue is discussed.

The outline of the paper is as follows: a consideration of SDBC; an analysis of the applicability of SDBC for building software brokers in the cultural heritage context; conclusions.

## The SDBC Approach

The SDBC approach is intended for supporting the alignment between business process modeling and software specification, within the software development context. Theoretical analyses, case studies, and expert opinions [Shishkov & Dietz, 2004-2] support the claim that SDBC provides appropriate mechanisms for the support of the specification task, concerning actual domains such as e-Business, Tele-Work, Virtual Organizations, and so on. The main contribution of the approach is the provision of the essential alignment between a model of the business system to be supported by software and the specification of the (software) system-to-be. Next to that, the derived specification model would be consistent with the current software standards, such as UML [OMG, 2004] and XML [XML, 2004]. Moreover, in such a derivation, SDBC allows for re-use of modeling components. The above-mentioned advantages of SDBC, which distinguish it from the currently used popular software design methods (such as Kobra, Catalysis, Tropos, and so on) [Shishkov & Dietz, 2004-1], relate to a number of issues among which is the adoption of the component-based modeling and specification. The component-based perspective has been addressed in the previous section.

Due to the limited scope of this paper and also because of the availability of relevant sources [Shishkov & Dietz, 2004-1,-2], the SDBC approach will not be introduced. Instead, a brief outline of some essential SDBC issues is presented below, supported by Figure 1.

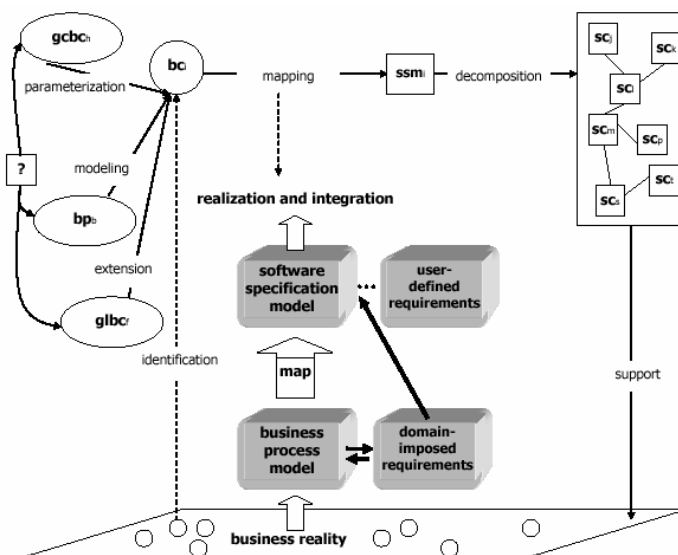


Figure 1. The SDBC approach: basic issues

The following abbreviations were used on the Figure:

bp = business process

bc = business component

glbc = general business component

gcbc = generic business component

ssm = software specification model

sc = software component

Due to the limited scope of this paper, definitions are not presented; interested readers could find them in the SDBC materials which were already mentioned.

As seen from the Figure 1, any business reality is viewed as composed of business components (related to corresponding business processes). Hence, an essential issue is the identification of business components. The identification mechanism is introduced in [Shishkov & Dietz, 2004-1]. According to it, a business component could be identified either by modeling of a business process (the trivial way) or by re-using 'pre-fabricated' components. SDBC distinguishes between two types of such re-usable patterns, namely general business components and generic business components. A general business component is a model in which some core functionality is grasped (for example, a general reservation unit). Hence, in order to derive a business component, based on a general one, it is necessary to apply extension (for example, extending a general reservation model into a hotel booking model). As for generic business components – they have in themselves several functionalities captured. Any of them could be selected as an option through parameterization (for example, if there is an accounting system, it might have two options: to work on the basis of the British law or to work on the basis of the French law; therefore, a user could adjust the system and use either of these options). Summarizing, a business component could be identified either by modeling of a business process, or by extending a general business component, or by parameterizing a generic business component. Since within SDBC, a business component is a model of a part of a business system, a crucial question is what are the demands towards such a model. As explained in [Shishkov & Dietz, 2004-1], the model should be soundly elaborated in the following three perspectives: structural perspective (concerning transactions, corresponding actors and their interrelations), dynamic perspective (concerning the transactions workflow representation), and communicative perspective (concerning the inter-actor exchange of communicative act which accompany the actual execution of transactions and are, therefore, of significant importance in specifying a software system if the goal is to adequately integrate such a system in the original business environment). Based on analyses of a number of relevant modeling techniques, DEMO [Dietz, 1999] has been selected for the purpose of facilitating SDBC as far as business components are concerned. The basic reason is the allowance (provided by DEMO) to adequately consider each of the above mentioned three perspectives. Hence, within SDBC, the identified business components are modeled using the notations of DEMO.

After being identified, a business component is to be mapped towards a software specification model. At this stage, the requirements related questions are to be considered. The domain-imposed requirements (those concerning the original business model) are to be derived based on the business component. The user-defined requirements (those especially formulated by the future users of the system-to-be) are added when specifying the software model. As for the specification model itself, it should be fully consistent with the current software design standards, as mentioned before. Hence, it is aimed that SDBC provides a UML-based software specification output. This means that the software specification model derived from a business component should be built with the use case notations; as it is well known, use cases are the modeling constructs serving to link the application domain (the business world) and the software domain, in a UML-based design of software. Thus, taking into account that business components are to be built using DEMO and that the resulting software specification model is to be built with the use case notations, a DEMO – use case mapping is an essential problem within SDBC. This particular problem has been addressed in [Shishkov & Dietz, 2003].

Further on, once derived, the software specification model is to undergo decomposition in order to give way to the identification of software components based on which the software system-to-be should be constructed. The reason for considering components also at this stage of the software creation process is related to the re-use possibilities which result from the component-based way of modeling and designing. Hence, by 'software component' is to be meant a (re-usable) part of a software specification model. Thus, the software components in SDBC are not physical components (associated with the current 'physical' component technologies, such as CORBA, .NET, J2EE/EJB, and so on), but rather 'logical' components representing the logical building blocks of a software system. It should be noted also that SDBC addresses just the business/software alignment and therefore goes as deep as software specification, in considering the software system-to-be.

Software components, specified in this way, could be further integrated and implemented using the above mentioned component technologies. The elaboration level of an SDBC component should correspond to the UML practices. This means that after producing a use case specification of a software component, the use case model needs further detailization (stakeholders analysis, main success scenario and extensions, triggers, and so on) and elaboration (it should include additional modeling activities aiming at complementing the use case diagram with other essential UML ones, such as the UML sequence diagram, the UML class diagram, and so on). Such elaboration is claimed to be sufficient for providing the further integration and implementation phases with an adequate modeling input. After a number of components are integrated and implemented, a resulting software application would appear. According to Figure 1, the essential goal of this application should be the support of the original business system.

For more information on the SDBC approach, interested readers are referred to the SDBC sources, mentioned before.

### SDBC, Brokers and Cultural Heritage Aspect

The SDBC approach is claimed to offer useful advantages concerning the specification of software systems that are intended to support complex business systems in different domains. Among the domains where SDBC has successfully been applied are e-Business and Tele-Work, as already mentioned. In both of them, it has been demonstrated how (via SDBC) a software brokerage system could be specified. As it is well known, software brokers are of great interest currently because of their wide applicability resulting from their actual (brokerage) functionality. Software brokers usually facilitate: 1. the match-making of globally available information, 2. the management of digital archives, 3. the globalization of used data networks. Through software brokers, users could have a quick and effective match-making at low costs.

Because of the relevance of software brokers to some cultural heritage issues (as already mentioned) and also because of the appropriateness of SDBC with regard to the specification of software brokers, it is suggested that SDBS is applied in building cultural heritage sector brokers. They could:

- effectively handle the management and global distribution of metadata as well as of digitized cultural/scientific information;
- be usable on a global scale through the Internet.

Therefore, (SDBC-based) software brokers could stimulate the global availability of cultural/scientific data. Below, the role of SDBC in building such brokers will be illustrated. Because of the limited scope of this paper, the illustration will be incomplete. Only some essential issues will be discussed. Starting from the original user information, the first task would be to understand correctly the business problem to be solved by the software system-to-be and to structure somehow the user information which is usually vague and unstructured. A simplified brokerage model has been drawn (Figure 2) around which the particular case information could be added and analyzed.

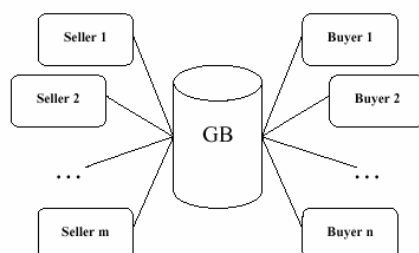


Figure 2. General brokerage functionality (simplified view)

As seen from the Figure, in the general case, there are a number of 'sellers' (distributors) of anything and a number of 'buyers' interested in it. The 'General Broker' (GB) should match appropriate seller and buyer information based on some criteria.

Using this general model, further analysis should follow, considering the particular cultural heritage information (briefly discussed already). In SDBC, the result of such an analysis is reflected in the so called 'SCI Model'. SCI

stands for Structuring Customer’s Information. Interested readers could read about the model in the above mentioned SDBC sources.

The general SCI model relevant to this particular situation is depicted (just for illustrative purpose) in Figure 3. The model is incomplete, only some basic issues are there. Among them:

- Units within the General Broker (GB):
  - = AU (Acceptance Unit): responsible for accepting and handling submissions from sellers and buyers;
  - = FU (Financial Unit): responsible for handling the fee payments done by sellers/buyers as a compensation for the work of the broker;
  - = MM (Match-Maker): responsible for performing the match-making concerning the seller and buyer information.
- Actors outside the General Broker:
  - = Seller (offering/distributing something, for example, digitized cultural materials);
  - = Buyer (being interested in something, for example, in particular digitized cultural materials);
  - = Expert: responsible for assisting the broker in some complex situations (in which, for example, real ‘human’ cultural heritage experience is required);
  - = Insurer: responsible for the insurance of relevant issues (for example, insurance against fraud).

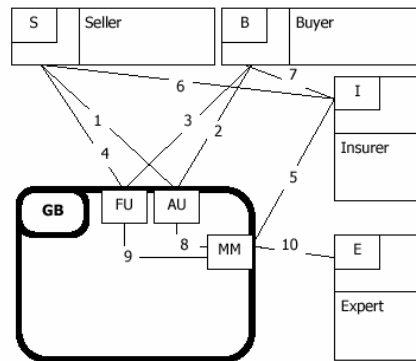


Figure 3: The General Broker: SCI Model

Hence, the GB SCI model facilitates the structuring of the initial case information to be reflected in the identification of a business component. In this particular case, it is suggested that a general business component is identified. The reason relates to the wide usage of software brokers in a number of cases; thus, identifying a general model would allow for re-using it many times. Our (DEMO-based) general business component is depicted in Figure 4. The model is incomplete because of its having just an illustrative purpose.

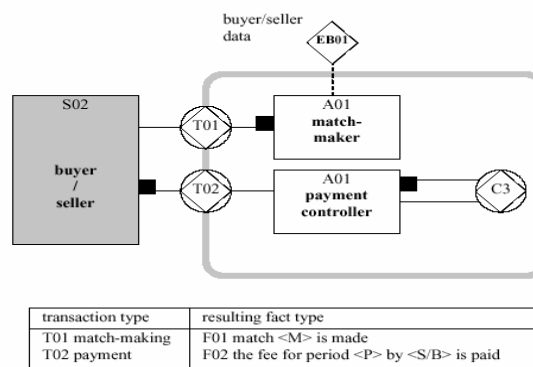


Figure 4: General business component (DEMO-based): The General Broker



As seen from the Figure, two internal GB units are depicted and also two external actors ('seller' and 'buyer': they are modeled as an aggregated actor because of their having the same general attitude towards the broker). As for the internal GB units depicted, they are the match-making unit ('match-maker') and the financial unit ('payment controller'). Two transactions are specified, concerning the actors and relevant GB units: T01 match-making (executed by the match-maker) and T02 payment (executed by seller/buyer). One transaction is specified, taking place within the GB: C3 (it concerns the periodical self-activation of the payment controller in handling all the payments related to a particular period of time). There is also a data bank depicted (EB01), containing the necessary data (concerning both buyers and sellers) that the match-maker should have in order to be able to realize a match-making.

Those readers not familiar with DEMO, are referred to [Dietz, 1999]. Once built, this general business component needs to be extended (as shown on Figure 1) aiming at the identification of a particular business component, in this case: Cultural heritage sector broker (built again with the DEMO notations). However, because of the limited scope of this paper, the transformation from the general business component (The General Broker) to the particular business component (The Cultural heritage sector Broker) is not presented. Information on how such an extension is carried out within SDBC could be found in the mentioned SDBC materials. Hence, a DEMO-based business component (The Cultural heritage sector Broker) should be reflected in a use case software specification model. An example of such a model is depicted in Figure 5. The model is incomplete, containing only some of the use cases characterizing such a broker. The broker is to use a database. It is virtually divided in two parts: one concerning the data submitted by distributors (of digitized cultural heritage materials) and the other one, concerning the data submitted by users. They are represented on the Figure by 'DBD' and 'DBU', respectively ('DB' standing for database; D(U) standing for distributor (user)). On the Figure, it is just illustrated how a software specification model would look like. The use case model will not be explained since it is expected that most of the readers are familiar with UML. As for the DEMO – use case derivation mechanism, information on it could be found in [Shishkov & Dietz, 2003].

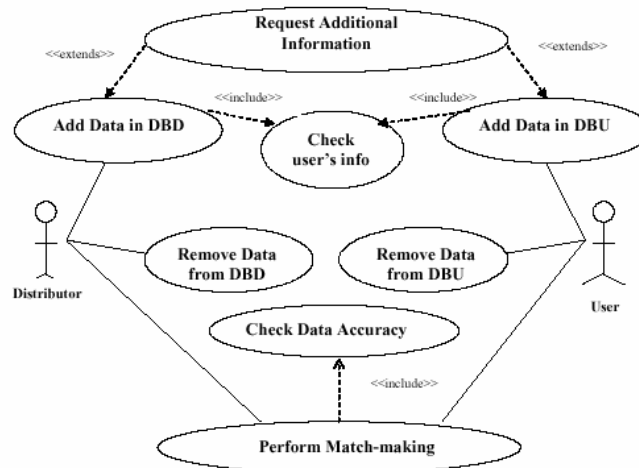


Figure 5. The Cultural Heritage Sector broker: Use case software specification model

## Conclusion

Being based on the innovative idea of aligning business process modeling and software specification in a component-based way, SDBC provides an appropriate framework for designing software systems allowing for a sound reflection (in the software model) of the original business requirements and usage of the relevant software standards. Therefore, using SDBC, software specification could be improved (in general) by:

- aligning it to prior business process modeling (based on (re-usable) business components);
- mapping precisely business process models to UML;
- applying powerful modeling techniques and environments, such as DEMO.

Besides its applicability in specifying software brokers in actual domains, such as e-Business and Tele-Work, SDBC proved to be useful in building cultural heritage sector brokers. The great complexity of the original business input makes this situation inadequately solvable by using the existing popular software design methods. The basic reason is that neither of them allows for a sound and complete business/software alignment. Hence, this would result in considering partially the initial business input and relying (as usually) on suitable interpretations from the software designers. However, in such a specific case this could not replace a necessary adequate consideration of the original business issues seen from the specific cultural heritage perspective. Thus, a business/software alignment is crucial and this motivates the selection of SDBC as a suitable approach with regard to the particularly considered (brokerage) problem. Brokers built using SDBC would:

- be rigorously rooted in the (original) business process model;
- be easily re-usable in other cultural heritage projects because of their being built based on components;
- be fully consistent with the current software design standards.

By supporting the specification of cultural heritage sector brokers, SDBC proves to be useful for the cultural heritage domain because such brokers could facilitate the management and global distribution of digitized cultural (and scientific) information.

---

## Bibliography

---

- [Dietz, 1999] J.L.G. Dietz. Understanding and Modeling Business processes with DEMO. In the Proceedings of the International Conference on Conceptual Modeling (ER'99), Paris, France, November, 1999.
- [Jacobson et al, 1992] I. Jacobson, M. Christenson, P. Jonsson, G. Overgaard. Object-Oriented Software Engineering: A Use Case Driven Approach, Addison-Wesley, 1992.
- [OMG, 2004] <http://www.uml.org>
- [Shishkov & Dietz, 2004-1] B.Shishkov and J.L.G.Dietz. Aligning Business process modeling and Software specification in a Component-based way, the Advantages of SDBC. In the Proceedings of the 6 th International Conference on Enterprise Information Systems (ICEIS'04), Porto, Portugal, April 14-17, 2004.
- [Shishkov & Dietz, 2004-2] B.Shishkov and J.L.G.Dietz. Design of Software applications using Generic business components. In the Proceedings of the 37 th Hawaii International Conference on System Sciences (HICSS'04), Big Island, Hawaii, USA, January 5-8, 2004.
- [Shishkov & Dietz, 2003] B.Shishkov and J.L.G.Dietz. Deriving Use cases from Business processes, the Advantages of DEMO. In: Enterprise Information Systems V. Ed. O. Camp, J.B.L. Filipe, S. Hammoudi, and M. Piattini. Kluwer Academic Publishers, Dordrecht/Boston/London, 2004.
- [Shishkov, 2002] B. Shishkov. Business Engineering Building Blocks. In the Proceedings of the 9 th Doctoral Consortium on Advanced Information Systems Engineering (CAiSE'02), Toronto, Ontario, Canada, May 27-28, 2002.
- [XML, 2004] <http://www.xml.org>

---

## Author Information

---

**Boris Shishkov** – Department of Software Technology, Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology; Mekelweg 4, 2628 CD Delft, The Netherlands;  
e-mail: [b.b.shishkov@ewi.tudelft.nl](mailto:b.b.shishkov@ewi.tudelft.nl)

---

## DIGITIZATION PROJECTS CARRIED OUT BY THE MATHEMATICAL INSTITUTE BELGRADE

Zoran Ognjanović and Žarco Mijajlović

*Abstract:* In this paper some current digitization projects carried out by the Mathematical Institute of Serbian Academy of Science and Arts Belgrade and the Faculty of Mathematics Belgrade are described. The projects concern developing of a virtual library of retro-digitized books and an Internet data base and presentation of electronic editions of some leading Serbian journals in science and arts, and the work on the South-Eastern European Digitization Initiative (SEEDI).

*Keywords:* digitization, cultural heritage, scientific heritage, virtual library, SEEDI

---

### Introduction

---

The Mathematical Institute of Serbian Academy of Science and Arts Belgrade and the Faculty of Mathematics Belgrade carried out some digitization projects in the last decade. The projects concern cultural and scientific heritage digitization as a very broad field. Some of the most important projects are:

1. PANDORA, an expert system for archaeology (1993-1994)
2. Computer archiving and multimedia presentation of cultural values and national heritage (1995-1996)
3. Archiving the journal Publications de l'Institut Mathématique (1995-) [5, 9]
4. Collected works of Bogdan Gavrilović (1996-2001)
5. Old maps, engravings and photographs – Collection of the City Museum of Belgrade (1996-1997, 2004) [6,8]
6. Presentation of Historical Archive of town Kotor (1996-1997)
7. Memorial compact disk of the Faculty of Mathematics in Belgrade (1998)
8. The work on the foundation of the National Center for Digitization (2002-) [16]

which are described in detail in [13].

In this paper we present some new projects of our institutions that are now in progress: developing of a virtual library of retro-digitized books and an Internet data base and presentation of electronic editions of some leading Serbian journals in science and arts, and the work on the South-Eastern European Digitization Initiative (SEEDI) [17].

---

### Virtual Library of Retro-digitized Books

---

Working on digitization of elderly works of Serbian authors in mathematical sciences was a part of some earlier projects, for example of the Collected works of Bogdan Gavrilović and the Memorial compact disk of the Faculty of Mathematics in Belgrade. Now, it is a project by itself led by Ž. Mijajlović. The aim is to create a comprehensive and interconnected collection of retro-digitized books and other digital documents written by our mathematicians or somehow connected to Serbia and Montenegro. The original works are mostly from the XIX century and the beginning of XX century. So far the following books and PhD-theses were digitized:

1. Boscovich Rogherio, Elementorum Universae Matheseos Tomus I, II, III, 1757
2. Dimitrije Danić, Konformno preslikavanje eliptičnog paraboloida na ravan, Jena, 1885.
3. Bogdan Gavrilović, Formiranje jednoznačnih analitičkih funkcija, Budapest, 1886.
4. M. J. Andonović, Kosmografija, Beograd, 1888.
5. Kosta Stojanović, Atomistika, Niš, 1892.
6. Đorđe Petković, Abelova teorema dokazana algebarski pomoću Riemann-ove teorije funkcija, Beč, 1893.

7. Mihailo Petrović, Sur les zéros et les infinis des intégrales des équations différentielles algébriques, Paris, 1894.
8. Petar Vukićević, Die Invarianten der Linearen Homogenen Differential-Gleichungen nTER Ordnung, Berlin, 1894.
9. Mijalko V. Ćirić, Racionalna Mekanika (Sveska prva), Beograd, 1897.
10. Bogdan Gavrilović, Teorija determinanata, Beograd, 1899.
11. Mladen Berić, Figurativni poligoni diferencijalnih jednačina prvog reda, Beograd, 1912.
12. Sima Marković, Opšta Riccati-eva jednačina prvog reda, Beograd, 1913.
13. Konstantin Orlov, Aritmetičke i analitičke primene matematičkih spektara, Beograd, 1934.
14. Đuro Kurepa, Ensembles ordonnés et ramifiés, Paris, 1935.
15. Milutin Milanković, Nebeska mehanika, Beograd, 1935.

At the moment a catalogue containing more than 500 items that belong to the libraries of the Faculty of mathematics and the Mathematical institute is under construction. We are also looking for the best way to present the digitized material at the Internet, which will ensure long-term durability and availability. One of the promising possibilities is to use the approach suggested by the National Library of the Czech Republic in the framework of the Memoria programme [12].

A more detailed description of this project is given in [14].

---

### **Internet Database and Presentation of Electronic Editions of Leading Serbian Journals in Science and Arts**

---

This project is cooperation between "Communication" [2] non-governmental organization and the Mathematical institute. A group of scientists and members of University of Belgrade formed "Communication" in order to promote understanding between people and nations in the region of former Yugoslavia and abroad. The idea was that a good starting point for improving communication between those nations with the same or very similar languages might be science and culture. According to that, it was decided to form an Internet database of freely accessible full-text journals [3] in those fields. The project leader is D. Grbić, while collaborators are Z. Ognjanović, V. Petrović, M. Korać and U. Midić.

Currently, data base contains 30 journals with more than 200 volumes, 3000 articles and 32000 pages. The presentations of the journals are dynamically generated from the database. Papers included in the data base could be searched (both in English and Serbian – by: authors' names, titles, titles of special sections within the journals, key words and words contained in abstracts), downloaded and printed.

The database and presentations were developed using Zope [20], an open source object-oriented environment for building web applications, under Linux operating system. Zope features a transactional object database, which can store not only content and custom data, but also dynamic HTML templates, scripts, a search engine, and relational database connections and code. One can use any Internet browser to read the presentation.

A part of the database and presentation contains almost 100 volumes of 5 mathematical journals published in Serbia:

1. Kragujevac Journal of Mathematics (4 volumes in English, 2000 -)
2. Matematički vesnik (22 volumes in English, 1993 -)
3. Nastava matematike (21 volumes in Serbian, 1992 -)
4. Publications de l'Institut Mathématique (42 volumes in English, 1980 -)
5. The teaching of mathematics (9 volumes in English, 1998 -)

The most recent project in this framework is called "Communication in Europe" [4]. Its aim is to make the database accessible in European alphabets and languages. A program package which enables translation of the interface and records was developed. It will be possible to enter in the data base the original or translated texts, abstracts, key words, titles, etc. in local languages and not only in Serbian and English. One of the expected long-term results of this project would be the standardization of domestic terminology and forming a thesaurus, including a comparative network of terms in European languages.

---

## South-Eastern European Digitization Initiative (SEEDI)

---

The South-Eastern European Digitization Initiative (SEEDI) [17] is an effort to develop awareness about digitization of cultural and scientific heritage in the South-Eastern European countries along the Lund Principles of the European Union [11]. It is expected that it will contribute to gathering and spreading specific and interdisciplinary knowledge from various institutions from the region and the European Union where leading experts in the field work.

The SEEDI arose from the cooperation between researchers from Belgrade and Sofia (prof. Dr Milena Dobreva from the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, prof. Dr Nikola Ikonov from Institute for Bulgarian Language of the Bulgarian Academy of Sciences) which was formally expressed in the Borovets declaration [1], a text concerning the future development of a network for digitization of scientific and cultural heritage in South-Eastern Europe which was formulated after the workshop "Digital Preservation of Cultural Heritage" [15] which was a part of the International Congress MASSEE'2003 [7]. In the declaration it was argued that »...[the researchers and the heritage institutions from the region] face common problems and share common scientific and cultural heritage. The knowledge and experience of single institutions from our countries should not stay isolated. It is of great importance to take measures for increase of the communication and exchange of technological expertise, standards and practical skills within the region, taking into account the experience of colleagues outside the region.«

It is well known that cultural and scientific heritage in our region is not widely available in electronic form. The idea of the SEEDI is to overcome that by bringing together researchers from regional and European centers having similar scientific and practical interest in digitization and to support cooperation between them. The aim is to create groups of specialists, which will be able to consult, assist, monitor and develop innovative technologies and digitization projects collaborating with the local heritage institutions.

The SEEDI would be implemented through several measures: conferences and workshops, the journal Review of the National Center for Digitization [10], the mailing list etc., to facilitate dissemination and sharing each other's ideas, concerns, views and experiences in the field. For example, the SEEDI-mailing list [18] is devoted to permanent communication between researchers involved in digitization of cultural and scientific heritage. It is planned that the First SEEDI Conference [19] will be held in Ohrid (Macedonia) under the name »Digital (Re-)Discovery Of Culture (Physicality Of Soul)« during the first week of September 2005 with the following topics: dance, music, playing (edutainment) and manuscripts.

---

## Bibliography

---

- [1] Borovets Declaration, <<http://www.ncd.matf.bg.ac.yu/?page=news&lang=en&file=declaration.htm>> [Date of last access: 2004-10-15]
- [2] Communication website, <<http://www.komunikacija.org.yu/>> [Date of last access: 2004-10-15]
- [3] Communication, Internet presentation of Serbian scientific and cultural journals, <<http://www.komunikacija.org.yu/komunikacija/casopisi/index?stdlang=en>> [Date of last access: 2004-10-15]
- [4] Communication in Europe, <<http://www.sac.org.yu/komunikacija>> [Date of last access: 2004-10-15]
- [5] European Mathematical Information System – EMIS, <<http://www.emis.de/>> [Date of last access: 2004-10-15]
- [6] Groman's Photo Album 1876 – 1878, <http://www.ncd.matf.bg.ac.yu/projects/en/groman.html> [Date of last access: 2004-10-15]
- [7] International Congress MASSEE'2003, 15-21. 9. 2003, Borovets, Bulgaria, <<http://www.math.bas.bg/massee2003/index.html>> [Date of last access: 2004-10-15]
- [8] Internet presentation of the CD "Old maps, engravings and photographs – Collection of the City Museum of Belgrade", <<http://www.mi.sanu.ac.yu/muzej.beograd/>> [Date of last access: 2004-10-15]
- [9] Internet presentation of the journal "Publications de l'Institut Mathematique", Mathematical Institute, Belgrade, <<http://www.mi.sanu.ac.yu/>> [Date of last access: 2004-10-15]
- [10] Internet presentation of the journal Review of the National Center for Digitization, Faculty of Mathematics, Belgrade, <http://www.ncd.matf.bg.ac.yu/?page=publications&lang=en> [Date of last access: 2004-10-15]
- [11] Lund Principles of the European Union, <[http://www.cordis.lu/ist/directorate\\_e/digicult/lund\\_principles.htm](http://www.cordis.lu/ist/directorate_e/digicult/lund_principles.htm)> [Date of last access: 2004-10-15]
- [12] Memoria Project website, <http://www.memoria.cz/> [Date of last access: 2004-10-15]

- [13] Ž. Mijajlović, Z. Ognjanović, A survey of certain digitization projects in Serbia, Proceedings of the Symposium Digital Preservation of Cultural Heritage, 16-17 September 2003, Borovets, Bulgaria, Review of the National Center for Digitization 4, 52-61, 2004. <<http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/4/d007/document>> [Date of last access: 2004-10-15]
- [14] Ž. Mijajlović, Z. Ognjanović, N. Đorđević, T. Zečević, VIRTUAL LIBRARY – data base of textual data, Proceedings of the Third International Conference »New Technologies and Standards: Digitization of National Heritage 2004« Belgrade, Serbia and Montenegro, June, 3-5, 2004, to appear in Review of the National Center for Digitization.
- [15] Minisymposium "Digital Preservation of Cultural Heritage", International Congress MASSEE'2003, 16-17. 9. 2003, Borovets, Bulgaria, <[http://www.math.bas.bg/massee2003/BAL\\_conference.html](http://www.math.bas.bg/massee2003/BAL_conference.html)> [Date of last access: 2004-10-15]
- [16] National center for digitization website, <<http://www.ncd.matf.bg.ac.yu>> [Date of last access: 2004-10-15]
- [17] The SEEDI website, <http://www.ncd.matf.bg.ac.yu/seedi> [Date of last access: 2004-10-15]
- [18] The SEEDI-mailing, [seedi@matf.bg.ac.yu](mailto:seedi@matf.bg.ac.yu)
- [19] The First SEEDI Conference »Digital (Re-)Discovery Of Culture (Physicality Of Soul)« website, <http://www.ncd.matf.bg.ac.yu/seedi/events/events.html> [Date of last access: 2004-10-15]
- [20] Zope website, <<http://www.zope.org/>> [Date of last access: 2004-10-15]
- 

### Authors' Information

---

**Zoran Ognjanović** – Mathematical Institute of Serbian Academy of Science and Arts, Kneza Mihaila 35, Belgrade, Serbia and Montenegro, email: [zorano@mi.sanu.ac.yu](mailto:zorano@mi.sanu.ac.yu)

**Žarko Mijajlović** – Faculty of Mathematics Belgrade, Srudentski Trg 16, Belgrade, Serbia and Montenegro, email: [zarkom@eunet.yu](mailto:zarkom@eunet.yu)

## MALTESE EXPERIENCE WITH DIGITIZING CULTURAL HERITAGE

**Charles Farrugia**

*Abstract.* The article gives an account of the various microfilming initiatives taken in Malta during the last thirty years. Various archives have managed to microfilm their holdings under co-operation agreements with international societies, or manuscript libraries. The advent of digital technology is now posing new challenges and opportunities for the archives sector. The idea of a National Memory Project that will try to bridge the different approaches in the preservation of records in the various public, private, and ecclesiastical archives in Malta is discussed. Technical challenges are highlighted, as are the opportunities that arise from collaboration and active participation in international projects such as the European Visual Archives (EVA), and the SEEDI initiative.

*Keywords.* Archives, Audio-Visual Archives, Cultural Heritage, Digitization, Malta, National Archives of Malta National Memory Project

---

### Introduction

---

Malta is a small island (450 km<sup>2</sup>) located in a very central position in the Mediterranean Sea. Its population does not exceed 400,000, yet its territory is very rich in cultural heritage. If one considers that no less than five pre-historic sites are situated within such a small territory, one appreciates the concentration of heritage on such a small island. These sites were declared world heritage sites by UNESCO. This paper does not attempt to delve into the diverse areas of cultural heritage, but restricts itself to archives. The perspective of this study is that of an archivist, greatly influenced by Malta's Euro-Mediterranean identity. It also brings in some of the experiences of a profession struggling to keep its identity in a country which formed part of the Commonwealth as from 1964, and the EU as from 1<sup>st</sup> May 2004.

---

## Maltese Archives

---

Archives provide the bedrock for our understanding of the past. They show us and future generations, how we came to be what we are as a nation, a community or an individual. They are a hidden national asset and constitute the very essence of our heritage. Malta is rich in archival holdings with Maltese notaries practising their profession and thus creating archival records way back into the 15<sup>th</sup> century. Archives in Malta are administered under different authorities and governing bodies, depending on whether they are public, ecclesiastical, or private. The main public archives i.e. those records created, maintained or received by institutions performing a public function, are: the National Archives, Notarial Archives, Public Registry, National Library (Archives Section), Department of Information (DOI) and Public Broadcasting Services (PBS).

Ecclesiastical archives are repositories administered either by the Diocese of Malta, or by the Cathedral of Mdina, or by any other religious order. The main archives are those of the Archiepiscopal Archives in Floriana, the Cathedral Archives in Mdina, Parish Archives, Religious orders, and the Wignacourt Museum.

Under the term private archives, we group together all those institutions whose functions are private, or whose shareholding in their administration is private. The main Maltese private institutions holding archives of national significance are the Strickland Foundation, Social Action Movement, Commercial Banks, Political Parties, Chamber of Commerce and the Times of Malta.

The main challenge is how to overcome the fragmentation of the sector with archival records under the responsibility of the National Archives (1530), the Notarial Archives (1465), the National Library (1107), the Public Registry (1863), the Law Courts (1900), Church Archives (c. 15<sup>th</sup> Century), and Private Archives (1296) [*Dates in brackets represent the approximate dates of the first documents held in a particular archive.*]. In order to bring together archival cultural heritage into a sort of one-stop shop for the general public, the idea of creating a virtual archives started gaining ground. The ultimate aim is to provide the Maltese/European public with easier access to the richness of Maltese archives.

---

## National Register of Archives

---

In order to facilitate research, the idea to compile a National Register of Archives was repetitively put forward by a number of experts. This is a relatively inexpensive and effective way of putting together all available catalogues on line on one portal. This would facilitate access to the public and guarantee more security to the records. This recommendation was made by Ann Williams in 1971 during the conference 'Maltese History: What Future? [1] More than thirty years later, computer technology can facilitate such a task and eliminate the logistical problem of where to house such a register. In fact, it is encouraging that on 27 September 2004, a new Archives Bill was presented before the Maltese Parliament by the Minister of Education Hon. Dr. Louis Galea. One of the articles in the new Bill provides for the setting up of the National Register of Archives [2]. In view of the available technology, it is hoped that the National Register of Archives will also have links to extensive digitised holdings that can be viewed from the comfort of ones' own home at the press of a button. With this aim in mind, the National Archives started a long term project of creation of digital content for the eventual register. [*The first two projects were the digitisation of the first thirty years of documentation from the Consolato del Mare records (1697-1730), and the publishing of an audio-visual cd entitled Political Personalities in Malta 1800-2004.*]

---

## Digitising Holdings

---

The first consideration was the digitization of the holdings. Several holdings were already microfilmed in the early 1990s either by the Hill Monastic Library of Minnesota or the Genealogical Society of Utah. Tests and analysis on the feasibility of having digitization from microfilms were conducted by two Italian firms, Global Microfilm Digital S.R.L. and Datadisc.it. The end results showed that images of high quality were possible in the case of microfilms that were preserved in adequate atmospheric conditions. [*The recommended conditions in line with the BS5454:2000 are almost impossible to apply in Malta (even after making the adjustments due to the geographical variations) without controlled environmental chambers for the storage of microfilms.*]The same cannot be said to parts of collections from the Cathedral Museum, where the low quality of chemicals used at the production stage, coupled with the inadequate storage conditions in which they were stored, led to early symptoms of vinegar syndrome and eventual disintegration of the films.

A proposal for the automated identification of Medieval texts has also been made by MaltaLinks, a commercial

enterprise researching into software applications for analysis of medieval texts. Tests have been made using records from 16th century notarial archives and 19th century National Archives records.

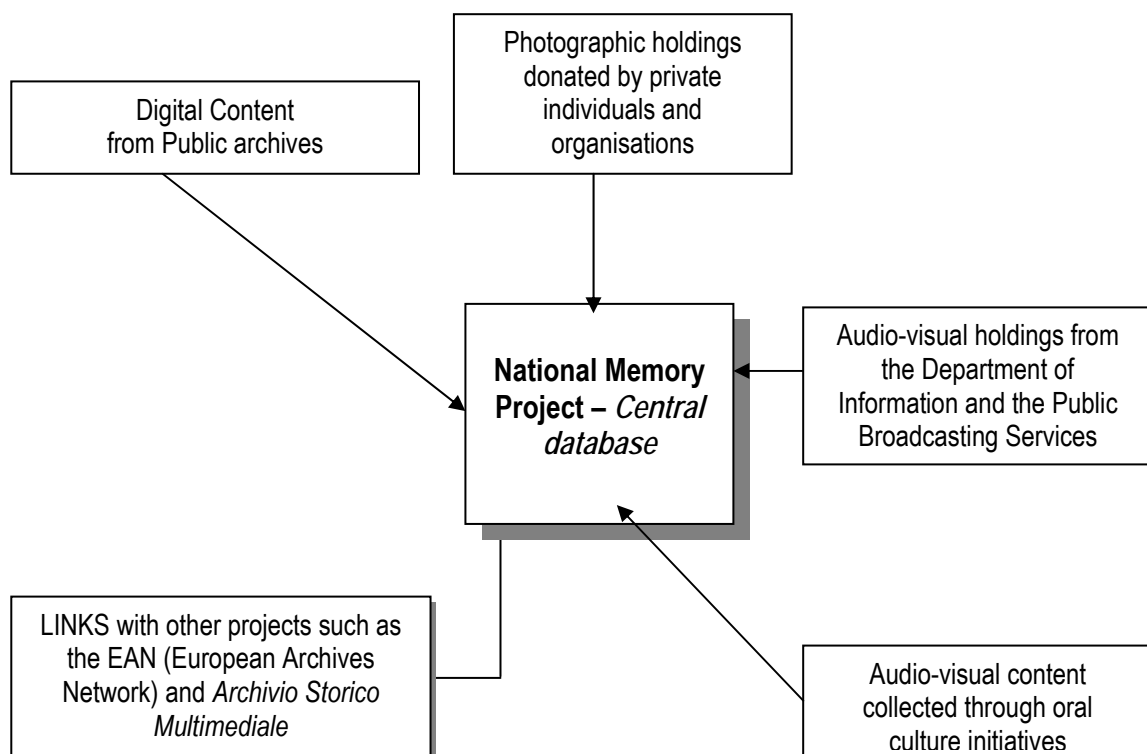
An archives of considerable importance for the study of Maltese toponymy and legal rights is that of the Notarial Archives. With the oldest documents going back to the 1460s, the archives provides the reading public with notarial deeds, wills, property transfers, loans, power of attorney etc. The cultural value of these holdings started being appreciated when a poem dating to the 1530s known as the Cantilena written by Pietro Caxaro was found on one of the end pages. The volumes form part of the notarial deeds of Notary Don Brandano Caxaro. A considerable part of the collection is already microfilmed and the conversion to digital will mean the creation of a preservation surrogates, the enhancement of accessibility to the holdings from the National Archives and also the future possibility of providing the contents via web-technology.

Several religious orders who are still active in Malta own extensive archival holdings of great historical and cultural value. The Augustinians, Dominicans, Franciscan Conventuals, Franciscan Minors, Capuchins, Jesuits and the various monasteries, preserve their records, some of which date back to the 14th and 15th centuries. Notwithstanding, accessibility is hampered and the reading public cannot appreciate the cultural value of such holdings. It is hoped, that through the national drive to digitise cultural heritage, these records are captured into a centralised system of access. An experiment was done last year with the digitisation of a volume from the Franciscans Conventuals archives of Rabat. A Guiliiana Antica dated 1638 was digitised and it is hoped that the authorities will appreciate the advantages of this joined national effort to create digital content and hence providing it to the international academic community.

Another active participant in the attempts to digitize Malta's cultural heritage was made by the University of Malta. The archives section of the University library holds the papers of prominent Maltese academics. As the first pilot project in digitization, they opted for the compilation on cd of all the papers of Fr. Karm Psaila, Malta's national poet. The cd includes an audio recording of the poet himself reading three of his poems.

### National Memory Project

The National Archives of Malta is the country's leading institution obliged under the 1990 Archives Act to preserve the archival heritage of the country. In order to make best use of digital technology to fulfil its mission, the National Archives of Malta has just launched an initiative called the National Memory Project (NMP). It aims to collect, preserve, and provide to the general public, Malta's photographic and audio-visual archives. It tries to bring together all stakeholders who are responsible or involved in the upkeep of cultural heritage in Malta and is also open for any initiative of collaboration with institutions from abroad.





---

The first phase of the project, which was inaugurated by the then President of Malta H.E. Professor Guido de Marco on 22nd March 2004, is the National Portrait Archive. It brings together portraits from archival records, portraits by prominent Maltese photographers and portraits donated by individuals, groups or organizations. A series of photographs of particular interest are the passport applications photos which total to around 100,000. Emigration dominated Malta's socio-economic history throughout the 19th and 20th century. The images captured with passport applications starting in 1915 provide ethnographic and historical detail of great value. All applications have been microfilmed by the Genealogical Society of Utah and plans are in the pipe-line to provide access to the images via the web in the near future.

The second phase of the NMP is the digitisation, cataloguing and intellectual control of thousands of historical photos. A public request was made for individuals and organisations to donate photographic material of historical value to the National Archives. In cases where private individuals wanted to keep the originals, scans were made and the bibliographical details recorded. Cataloguing fields prescribed by the International Standard for Archival Description (ISAD(G) and other standards for the cataloguing of photographs were used. It is hoped, that during 2005, the demo of this project will be available to the general public and that parts of the holdings will also be integrated with the European Visual Archives (EVA) initiative.

The third and final phase of the NMP will be the setting up of the Film and Sound Archive. Thousands of films in 16mm or 35mm are deposited in repositories ill-fit for such a purpose. The best available technology has to be identified to convert the images on durable and readable formats. Cataloguing has to be done in line with modern practices. This is no easy task as the cost involved is considerable. However, a small state such as Malta cannot afford to loose the audio-visual heritage it has generated throughout the last century. It represents its transition from a fortress colony to an Independent Republic. The images in question record Malta's salient socio-political events, the building of tourism as Malta's main economic base and depict the way the Maltese lifestyle developed and changed to its present state.

---

## Conclusion

Digitising Maltese cultural heritage is not an easy task. It was only lately that new structures were set-up to manage effectively such heritage. The enactment of the Cultural Heritage Act, the setting up of the Superintendence of Cultural Heritage, and Heritage Malta, as the agency responsible for the management of Malta's heritage brought with them major changes. The archives sector will soon pass through a similar reform. With the legal structures in place, and an ever-growing awareness amongst the general public, the prospects of a quality leap forward are very good. Malta's accession to the European Union has facilitated to movement of expertise and personnel. It is hoped, that digital technology will be used to its utmost to promote the country's heritage treasures. In an ever-growing globalised economy, Malta needs to promote its cultural heritage in the most attractive and widely accessible manner. It is hoped that collaboration with projects such as SEEDI, and European and Mediterranean initiatives, make it possible to share expertise, make best use of human resources, and facilitate the paths to be followed in order to avoid duplication of efforts, and create centres of information that can be easily accessible throughout Europe, and world-wide.

---

## Bibliography

- [1] Vella Bonavita, R. u Williams, A. (1974), Maltese history What Future?  
[2] Text of Bill available on-line on [www.parliament.gov.mt](http://www.parliament.gov.mt).

---

## Author Information

Charles Farrugia – Head of the National Archives of Malta. E-mail: [charles.j.farrugia@gov.mt](mailto:charles.j.farrugia@gov.mt)

## *PRESENTATIONS OF BULGARIAN INSTITUTIONS*

### **NATIONAL ARCHIVES**

**Nikolay Markov**

*Abstract: Digitisation as an opportunity for Bulgarian National Archives to fulfill its functions as a heritage Institution and a government agency according to the new standards.*

*Keywords: Bulgarian National Archives, information technologies development, digitization strategy*

---

#### **Information Environment and Functions of the National Archives**

---

The influence of the information technologies development on the institutional policy and current practices of the Bulgarian National Archives is determined by several factors:

- specific role of the Archives as a heritage institution on the one hand and as a government agency on the other;
- character and amount of users' requests;
- value, nature, physical condition and size of holdings;
- technological, financial and personal resources available to the Institution;
- availability of computer equipment to different kind of users;
- the laws in force.

The centralized system of Bulgarian National Archives was founded in 1951. It is composed of Central State Archives, Central Military Archives and 27 Regional State Archives all of them under the authority of the General Department of Archives at the Council of Ministers. Since 1960 the Institution has been member of the International Council on Archives.

In general, the holdings encompass records created by institutions, organizations, economical, political and cultural bodies and persons in the period from the Liberation of Bulgaria in 1878 till nowadays. Manuscripts from the age of Ottoman empire are preserved as well. Traditional textual records based on paper prevail, but we hold also architectural blueprints, engineering documents, photographic materials, cartographic archives, records on leather, motion pictures, audiotapes. The total size of collection is approximately 61000 linear meters.

As an archival heritage organization the National Archives has statutory functions of acquisition, preservation and use of archival records. By these functions it facilitates the inclusion of the historical records into contemporary cultural context.

On the other hand, as an Institution responsible for the management of the State archival fonds, the Archives exercises jurisdiction over provenances' records management and preservation and play a decisive role in identifying the records which will be part of the archival holdings in the future. Therefore, as the amount of records in electronic form produced in provenances' current practices increases, the responsibility of the Archives to establish standards for their management and transfer becomes obvious.

In the new information environment the traditional forms of reference service proved to be unsatisfactory to meet the new standards for access to relevant information. Outreach programs in conventional form as documentary publications based on paper, documentary exhibitions, motion pictures are also insufficient to attract wider audience and to encourage use of archival records. The researchers' requirements for faster and easier access increase; the circle of potential users with access to the Internet is wider; new standards for efficiency of public services emerge; necessity of new forms of collaboration between heritage organizations becomes evident. On the other hand we are obliged to prevent further deterioration of damaged or endangered originals without restricting the users' access to their content.

---

Digital technology is necessary for Bulgarian National Archives to accomplish its engagements to the establishment of electronic Government. According to the official government program we are obliged to provide access to records stored on electronic medium through the Internet. Archival records in our stacks are based on paper for the most part, but as a public administration we will be obliged to provide online access to some of them regardless of the medium. In addition, as mentioned above, the number of records produced in or converted to electronic form increases and we have to be ready for their acquisition, preservation and communication. To meet these challenges we have to elaborate a digitisation strategy.

---

### **Current Situation**

---

Today we are at the very beginning on this way. We have no long-term digitisation program or defined policy. Our first occasional attempts have been made mainly for the purposes of popularization. In 2003 Central Military Archives published on CD a documentary collection "The independence of Bulgaria and Bulgarian army". Now General Department of Archives intends to digitise the archives of the former Bulgarian communist party. The success of this first integral project will be of great importance for our future digitisation enterprises.

---

### **Perspectives**

---

Several opportunities offered by digital technology are especially attractive for us:

#### **1. Digitisation of high demand and most valued records.**

We could produce digital copies of certain archival groups and records according to criteria of archival value, high use and physical condition and make them available on CD or through the Internet according to the laws in force.

#### **2. Digitisation on demand.**

Another form is to provide the researchers with online access to digital images on individual requests. After loading these images, they will be available to all researchers under specified conditions. We could start this on the basis of conventional requests, but this service will be more effective after providing online information about our collection.

#### **3. Digitisation of records published in traditional form.**

We have produced many documentary publications and reference books in traditional form. Since the work of compilation and description of documents has already been completed, their content has been Word formatted and related finding aids have been produced, it will not be difficult to transform some of these publications into electronic form on CD. They will consist of scanned images of documents and related texts in appropriate format. Thus we will combine the possibility to view the records in their original form with readability and searchability of their formatted content. This is especially important in the occasions where the original text is faded or illegible.

#### **4. Multimedia publications composed of digital images of records, explanatory texts, motion pictures, studies.**

---

### **Benefits**

---

The profits provided by digital technology for the needs of access to archives are indubitable. Conversion of records in digital format is important also for the purposes of archival preservation. Although digitised copies should not be considered as the only conversion form for long-term preservation due to problems of "future proofing" or compatibility and authenticity, they allow to reduce handling of threatened originals. Possibility to provide faster and easier access to relevant information in different form for different audience will encourage the researchers to use the copies instead of paper records. The potential of digital form for access could be combined with long-term preservation measures in a balanced strategy.

---

### **Problems**

---

To clear the way for successful digitisation projects we have to overcome some difficulties of technical, legal and financial character.

1. The text of many highly valuable historic manuscripts has faded and the contrast has weakened. To guarantee readability of such documents in digital form we have to face specific requirements for reproduction and description quality.

2. Due to the nature of paper archives some widespread systems as feeder scanners and flat bed scanners are unsuitable.

3. Rapid succession of innovations of hardware and software requires regular measures for the safe migration of digital conversion form to preserve readability of information in new system environments.

4. National Archives provides the users with authenticated copies of documents to be used as legal instruments. If we are expected to provide online access to them, we have to solve some problems of authenticity and privacy.

5. An effective digitisation program requires financial and technical resources not available to the Archives in isolation. If we intend to develop such program instead of occasional digitisation trials, we have to establish long-term partnership with other organizations concerned with the cultural heritage.

If we are to continue to fulfill our functions according to the new standards and to provide services relevant to the new information environment, we cannot afford to ignore the possibilities offered by digital technology.

---

### Author Information

---

**Nikolay Markov** – General Department of Archives at the Council of Ministers, Sofia State Archives, Head of Sector, 5 “Moskovska” St., Sofia – 1000, BULGARIA; e-mail: [gua@archives.government.bg](mailto:gua@archives.government.bg)

## THE ROLE OF THE NATIONAL LIBRARY IN PRESERVING NATIONAL WRITTEN HERITAGE

**Elissaveta Moussakova and Alexandra Dipchikova**

*Abstract:* The first part presented at the meeting by A. Dipchikova is a brief report of the role of the National library as an institution in collecting, preserving and making accessible the national written heritage. Problems of digitization are examined from the point of view of the existing experience in cataloguing. Special attention is paid to the history and the significance of international standards, the experience in the field of development and maintenance of authority files on national and international level as well as in markup languages. Possibilities of using MARC and XML in the library are discussed. The second part presented here by E. Moussakova is giving an overview of the latest activities of the Library in the sphere of digitisation of the old Slavic manuscripts which are component of the national cultural heritage. It is pointed out that the current work is rather limited within the scope of preparation of metadata than being focused on digital products.

*Keywords:* National Library, Slavic manuscripts, digitization, preservation

---

### The Role of the National Library in Preserving the National Written Heritage

---

The aim of my presentation is to draw attention to the SS. Cyril and Methodius National Library as an institution closely related to the problems debated here by virtue of its main tasks and functions. Generally when defining the typology of national libraries, they are seen as such as they are a depository of national publications, they are the treasury of the written memory of the country and because frequently they are the coordinators of common activities of various institutions, directed towards the preservation of the cultural heritage. The classical concept of the national library includes the basic functions of collecting the national written heritage (printed matter and manuscripts), its preservation for the future generations and provision of the widest access to it for its contemporaries.

In order to meet these responsibilities, the National Library, collects current editions under the provisions of the Legal Deposit Law. At the same time, the library works to enrich its collections of manuscripts, old printed books, archival units which reach the library through donations and new acquisitions. Through the national library the government carries out its responsibility for the preservation of the national written heritage.

However, the active preservation of the collections of the library and provision of access to them to the general public and researchers is not only connected with their physical conservation and service organization, but also through the building up of bibliographic information and its dissemination.

The SS. Cyril and Methodius National Library has 125 years history. Throughout this period it has been the

leading institution entrusted with the preservation of the written heritage of the Bulgarian nation. The collections are organized and kept in the Archive of Bulgarian Literature. The SS. Cyril and Methodius National Library is the largest depository of manuscripts and old printed books in the country, while the Bulgarian Historical Archive keeps documents connected with the history and culture of Bulgaria from the Revival period to the First World War. The National Bibliography is prepared and published by the SS. Cyril and Methodius National Library.

In this way the library aims at assuming a leading role among other libraries and institutions engaged in the literary heritage. Today, after the decentralization of the library system of the country, this role is realized through functions with a multiplication effect on the system (for instance central cataloguing and the distribution of machine readable bibliographic records through the Internet) or participation in joint activities.

The absence of a common policy of successive governments in the field of librarianship has caused difficulties in determining Library's own strategy. The SS. Cyril and Methodius National Library has been forced to maintain its principal activities only under the conditions of a defined and limited budget. The outcome of this situation has for the time being been the preparation of projects funded partially or entirely by NGOs or international programmes. Over the last years the National Library has prepared several projects for the creation of electronic bibliographic information. This is the place to mention Bulgarian books 1878-1992 (a bibliographic database constituting the national bibliography over 115 years), the Bibliographical database of Bulgarian old printed books 1806-1878 project and the Authority files for the names of Bulgarian authors 1878.

These projects, some of which are completed and others in the course of completion show that the library has acquired some experience in the creation and organization of metadata. In the development of its bibliographic and cataloguing practice over the years, the National Library has followed the existing international standards. The problems of digitalization which were debated at this forum are frequently close to this practice and we believe it would be possible in future some of them to have an analogical solution.

This is why I would like to outline several fields where the experience of cataloguers in libraries could be useful.

Above all, I would like to mention the successful practice of international coordination of rules and standards in the field of bibliographical description which has been practiced for almost half a century. The program for Universal Bibliographic Control, devised and disseminated by UNESCO envisages a wide exchange of metadata of publishers' production from various countries and is the basis of continuing efforts in this respect. The increasingly successful attempts for coordination in exchange formats for machine readable bibliographic records are in the same line. The problems of the presentation of given names in bibliographic databases have an international significance. The national libraries of the separate countries, together with the Bulgarian National Library are working on special databases presenting the names of national authors in all their variety and in this way facilitate the global use of databases through the net. The long use of the use of the MARC markup language and the recent possibilities for the transfer of MARC to XML provide conditions for the coordination of an enormous amount of existing bibliographic information with metadata necessary in the process of digitalization of documents on paper carriers.

These examples show the necessity of processes of digitalization especially in the realization of projects on a national scale to be coordinated not only from the point of view of organization but also taking into account the experience acquired in related activities.

---

### **Special Collections in the National Library, Sofia**

---

#### *The Slavic Collection of Manuscripts*

The special collections at SS. Cyril and Methodius National Library, making a specific component of the national cultural heritage, are organized mostly within the Center for Manuscripts and Documents at the Library. These are the Slavic, Greek (plus other European languages), and Oriental (Arabic, Persian, Turkish) manuscripts, Slavic Old Printed books, Bulgarian Old Printed Books (1806–1878), rare and precious books, Oriental printed books, Ottoman archives, Bulgarian historical archive, and a rich collection of old photographs. Other collections like musical editions, maps and graphics are in charge of a separate department.

The National Library represents the largest repository of Slavic manuscripts in Bulgaria. It comprises about 1500 items dating from between eleventh and nineteenth centuries, most of which are already registered in five descriptive catalogues [Conev, 1910; Conev, 1923; Stojanov, Kodov, 1964; Stojanov, Kodov, 1971; Hristova, Karadzova, Vutova, 1996]. A Union catalogue of Bulgarian manuscripts from the eleventh to the eighteenth

century kept in Bulgaria has been published in 1982 by a team of Library researchers [Hristova, Karadzova, Ikonomova, 1982].

The focus of my presentation is the Slavic manuscript collection managed by the Department of Manuscripts and Old Printed books. In a nutshell, our activities in digitisation of manuscript heritage are still limited within the sphere of preparation of metadata. It means that our efforts are mainly spent on discussing, giving expertise and participating in projects related with electronic descriptions of medieval manuscripts rather than on producing digital editions. However, some experience in digitising has been acquired, with the CD "Unique Balkan Manuscripts" issued in 1994, on which images of Gospels, Gospel Lectionaries and Qur'ans from the Library have been put. The total number is 1030 images taken from Greek, Bulgarian, Serbian, Walachian and Moldavian, and Oriental manuscripts, each supplied with short description. While the selection aimed at showing to a broader audience interesting samples of medieval illustration, illumination and art of writing, some of the reproduced texts were chosen by certain symptomatic features as to make the electronic tool useful for textological studies. For its time this was the first product of its kind in the country and the quality of the images was very satisfactory. A recent product of the Library is the "Guide to the Special Collections in the National Library", a CD edited last year as part of the program for celebrating the 125th Anniversary of the Library, in which pages of well-known manuscripts have been included.

---

### **The Projects**

---

More important and adequate to the purpose of this kick-off meeting is the question what is the Library's latest work in the sphere of presentation and preservation of its valuable Slavic manuscript collection. To answer it I would like to make you acquainted with our already realised projects, what I partly did, current projects, and projects in plan.

Among the finished projects it is worth to mention the participation of the Department of Manuscripts and Old Printed Books in two important initiatives. The first one was the Library project for Retroconversion of the Bulgarian Printed Books, of which the Retroconversion of the Old Printed Books became the conclusive element. With the financial support of the Open Society Foundation and using – and, actually developing and adapting – the standard ISBD (A) within ISIS, the process was completed last year. At the moment a full revision of the records is being carried out, to be followed by building the database and making it available through the website of the Library. The second project was started as cooperation between the Library and the Institute of Mathematics and Informatics at BAS with the aim to apply on Slavic material the standard for electronic descriptions of medieval manuscripts, offered by the international group MASTER. The task of the Library, recognised as associated member of the group, was to prepare test descriptions of manuscripts from our collection, done on the base of sixteen codices in three different modes: short description following the model of the Union catalogue, more developed description on the level demonstrated by the Descriptive catalogues, and finally, very detailed description not made so far by other specialists. Unfortunately, some circumstances beyond our responsibility prevented putting the records on the Internet so the results, even if reported at a meeting of MASTER in Copenhagen in 2000, remained unknown to the specialists. Approaching its conclusion is the international project under the auspices of the Department of Medieval Studies at the Central European University in Budapest and the National Széchényi Library, Hungary for describing the Slavic manuscripts kept there. The electronic tool on which the team agreed was the MASTER standard but for the purpose, the DTD had to be reworked since it proved its incompatibility with some peculiarities of the Slavic manuscripts and also with the tradition of the already printed catalogues [Cleminson 2003]. In order to furnish the catalogue with illustrations, one or more digital images were taken out of each manuscript.

On the border between the accomplished and still running, the Budapest project is similar in its final goal to a current project envisaging the electronic issue of the Union catalogue of Bulgarian manuscripts. In its essence it is a retroconversion which the National library planned to do in yet another cooperation with the Institute of Mathematics and Informatics, represented by the group lead by Dr. Milena Dobрева. The descriptions of the manuscripts will be in an electronic format developed from the MASTER standard and adapted to reflect the structure and content of the Union catalogue. Our expectations are that once finished, the catalogue will become available on Internet through the portal of the National Library. If successful, this will be the first electronically accessed catalogue of medieval manuscripts in Bulgaria.

Among the future projects I will rank first the Digital Enina Apostle, which is a project actually started and then

interrupted, due to some technical problems. Another project aiming at educational purposes, but still in gestation, is a Digital Paleographic Album of the Thirteenth Century Hands and Scripts documented by codices kept in the National Library.

An important aspect of the projects are the partnerships on which the Library relies for realisation of its plans. Here the Institute of Mathematics and Informatics, headed by Prof. Dr. Stefan Dodunekov, must be mentioned as our main partner, both institutions being bound by a signed agreement for cooperation. The duties are divided in such a way that the Institute undertakes the "technical" part, having the high quality equipment while the Library participates with the "intellectual" capacity of its cataloguers and paleographers. To effectuate the project of the Electronic Union Catalogue, originally a demand of the Director of the Library Prof. Dr. Borjana Hristova, the team of M. Dobreva is putting the data from the printed catalogue into the electronic format and the records are to be checked and revised by the employees of the Manuscript Department. A similar distribution of roles is suggested for the Digital Enina Apostle, an intriguing and provocative project of digitising the earliest Slavic codex extant in Bulgaria, which, even if restored, is in fragile condition. Another partnership, also institutionally supported, is the one with the Department of Old Bulgarian literature at the Institute of Literature, to whose project of creating an electronic repertory of Old Slavic texts the Manuscript Department is giving consultation mostly on issues concerning manuscript illumination and marginal notes. During the work on the Széchenyi catalogue a successful partnership was established by the Hungarian and Bulgarian National Libraries of which the first undertook the digitisation of the folios selected from the manuscripts.

Whatever future activity in the domain of digitisation of old manuscripts is envisaged, it is always determined by two priorities of the Library – preservation of and access to its special collections. Therefore a prospective thinking ought to consider first of all the endangered volumes and fragments. Related to the issue is a question, perhaps a minor one, which I shall leave open not only before the present audience but perhaps to the librarians to come to the Library: what to do with the incomplete and low quality microfilm stock? Should we microfilm anew the whole collection or should we, in a long term project, digitize each manuscript, to answer the demands of the modern information society?

---

## Bibliography

- Cleminson 2003: R. Cleminson. A Good Servant but a Bad MASTER: Uses and Abuses of Standards in Manuscript Description. In: Computational Approaches to the Study of Early and Modern Slavic Languages and Texts. Proceedings of the "Electronic Description and Edition of Slavic Sources" conference, 24–26 September 2002, Pomorie, Bulgaria. "Boyan Penev" Publishing Center, Institute of Literature, Sofia, 2003, pp. 105–111.
- Cleminson, Moussakova, Vutova 2003: R. Cleminson, E. Moussakova and N. Vutova. Description of the Slavonic Cyrillic Codices of the National Széchenyi Library. In Annual of Medieval Studies at CEU, Vol. 9, 2003, pp. 339–348.
- Conev, 1910: B. Conev. Opis na rākopisite i staropechatnite knigi na Narodnata biblioteka v Sofija. T. 1. Narodna biblioteka, Sofija, 1910.
- Conev, 1923: B. Conev. Opis na slavijanskite rākopisi v Sofijskata narodna biblioteka. T. 2. Narodna biblioteka, Sofija, 1923.
- Hristova, Karadzhoa, Ikonomova, 1982: B. Hristova, D. Karadzhoa, A. Ikonomova. Bālgarski rākopisi ot XI do XVIII vek zapazeni v Bālgarija. Svoden katalog. T. 1. Narodna biblioteka "Kiril i Metodij", Bālgarska arheografska komisija, Sofija, 1982.
- Hristova, Karadzhoa, Vutova, 1994: B. Hristova, D. Karadzhoa, N. Vutova. Opis na slavijanskite rākopisi v Sofijskata narodna biblioteka = Catalogus manuscriptorum slavica quae in Bibliotheca Serdicensi asservantur. T. 5. Narodna biblioteka "Sv. Sv. Kiril i Metodij", Bālgarska arheografska komisija, Sofija, 1996.
- Kodov, 1983: Eninski apostol. Faksimilno izdanie s predgovor ot Hr. Kodov. Nauka i izkustvo, Sofija, 1983.
- Stojanov, Kodov, 1964: M. Stojanov, Hr. Kodov. Opis na slavijanskite rākopisi v Sofijskata narodna biblioteka = Catalogus manuscriptorum slavica quae in Bibliotheca Serdicensi asservantur. T. 3. Dārzhavno izdatelstvo Nauka i izkustvo, Sofia, 1964.
- Stojanov, Kodov, 1971: M. Stojanov, Hr. Kodov. Opis na slavijanskite rākopisi v Sofijskata narodna biblioteka = Catalogus manuscriptorum slavica quae in Bibliotheca Serdicensi asservantur. T. 4. Nauka i izkustvo, Sofia, 1971.

---

## Authors' Information

**Elissaveta Moussakova**, PhD – SS. Cyril and Methodius National Library, Head of the Manuscripts and Old Printed Books Department; 88 Vasil Levski Blvd., 1037 Sofia, Bulgaria; e-mail: [musakova@nationallibrary.bg](mailto:musakova@nationallibrary.bg)

**Alexandra Dipchikova**, PhD – Cataloguing Department in SS. Cyril and Methodius National Library; Head of the Scientific Board at the Library; 88 Vasil Levski Blvd., 1037 Sofia; e-mail: [dipchikova@nationallibrary.bg](mailto:dipchikova@nationallibrary.bg)



## INSTITUTE FOR BULGARIAN LANGUAGE, BAS

Vassil Rajnov

*Abstract.* The paper presents the history, structure and ongoing activities of the Institute for Bulgarian Language of Bulgarian Academy of Sciences.

*Keywords:* Bulgarian language, grammar, vocabulary, lexicology, lexicography, dialectology, etymology, onomastics, general and applied linguistics, corpora, phonetics, speech communication, computer modelling.

---

### Introduction

---

The Institute for Bulgarian Language (IBL) is the oldest Institute at the Bulgarian Academy of Sciences (BAS). It dates back to May the 15th 1942, when the Bulgarian Dictionary Office was founded at the Presidium of the BAS. IBL is the main national centre for study and description of the Bulgarian language – its present state, history, rich variety of dialects and relations with other languages. The Institute is a central and coordinating unit which defines the national language policies and establishes contacts with foreign institutions, interested in the Bulgarian language.

---

### Brief History

---

The founding of the Bulgarian Literary Society in 1869 marked the beginning of a new stage in the development of the Bulgarian philological studies. This event was called upon the necessity to assist the development and improvement of the Bulgarian literary language and to put on broad foundations the studying of the Bulgarian history and literature. In 1911 the Bulgarian Literary Society was restructured into Bulgarian Academy of Sciences. The new institution started its activity by establishing an improvised Dictionary Committee, which initiated the compilation of a dictionary of the Bulgarian language. Extensive lexicographic work in the following years led to the establishment of the Bulgarian Dictionary Service at the Bulgarian Academy of Sciences and Arts in 1942. After the re-organization of the Academy, the Bulgarian Dictionary Service was renamed as Institute for Bulgarian Dictionary in 1947 and as Institute for Bulgarian Language in 1949.

Nowadays the Institute for Bulgarian Language is one of the most respectable academic institutions with great contributions to education and language competence.

---

### Structure of IBL

---

The Institute consists of 9 departments, one laboratory and 3 services, which along with the theoretical language studies provide basic material for work in the field of applied linguistic. Further down a brief description of the activities of the particular divisions is provided.

The department for Modern Bulgarian Language carries out studies and supplies descriptions of the contemporary standard Bulgarian – its sound system, grammar and vocabulary. The theory and history of Modern Bulgarian are investigated, thus creating resources to enhance the linguistic competence of the society.

The department for Bulgarian Lexicology and Lexicography is the major national centre for compiling dictionaries and for training specialists in the fields of semantics, lexicology and lexicography. Its main task is the preparation of a multivolume academic explanatory Dictionary of the Bulgarian Language (DBL).

The department for History of Bulgarian Language works on the recovery and publishing with critical comments of old and mediaeval Bulgarian written records and carries out lexicological and lexicographical studies of the history of the Bulgarian language from the 9th to the 19th century.

The department for Bulgarian Dialectology and Linguistic Geography is a unique research section which carries out comprehensive studies of the language regional varieties on all levels. The linguo-geographical description of Bulgarian dialects is further used to the solution of some present-day problems of national significance.

The department for Bulgarian Etymology and Onomastics is engaged in compilation of the Bulgarian Etymological Dictionary, a fundamental publication of national significance, and the Dictionary of Bulgarian Toponyms.



The scope of the research in the department for Computer Modeling of Bulgarian Language includes: theoretical problems of the formal language description; formal semantic, morphological and syntactic analysis; information retrieval and information extraction; electronic dictionaries, and corpus linguistics.

The department for General and Applied Linguistics explores the theoretical and applied perspectives of general linguistics, sociolinguistics, anthropological linguistics, psycholinguistics, and the philosophy of language. Current research projects are dedicated to Europeanization trends, language manifestations of mentality, theoretical and applied problems of language communication, structural and social semiotics, ethnography of communication, etc.

The laboratory of Phonetics and Speech Communication carries out studies and practical work on restoration and digitization of audio archives. Another field of activities is dedicated to Speech Recognition and Text to Speech Conversion (TTS).

The department for Ethnolinguistic and Cultural Investigation of Bulgarian Language researches the major cultural phenomena of the Bulgarian tradition and the linguistic form and cultural semantics of these phenomena. The cultural study is focused on the ways in which Bulgarian cultural concepts and stereotypes reflect on the language as well as the forms and means of verbal communication, specific for the Bulgarian society of past and present days.

The department for Contrastive Investigation carries out research of the Bulgarian language in comparison and in contrast with other Slavonic, Balkan and European languages.

The Service for Bulgarian Terminology is the only specialized research unit in the country which studies Bulgarian terminology from a linguistic point of view and where terminology theories and various methodologies for practical work are developed. The Information Service and the Library are storing unique linguistics information. The Electronic Archive collects, updates, and processes Bulgarian electronic texts for the purpose of creation and maintenance of a representative national language corpus, bilingual and multilingual corpora.

---

### **Projects and Activities**

On the basis of abundant material – national archives and electronic corpora – various projects of national importance are implemented at IBL. These are mainly dictionaries of the Bulgarian language: multivolume explanatory, monolingual, bilingual, etymological, historical, spelling, phraseological, dialectological, topical, historical, corpus-based, etc., as well as atlases of Bulgarian dialects.

Beside the theoretical academic research one of the main objectives of the Institute for Bulgarian Language has been the creation of normative grammatical descriptions of the Bulgarian language. The three-volume Grammar of the Modern Bulgarian Literary Language is considered one of the most authoritative grammar descriptions of Bulgarian.

IBL is an active participant and reputable partner in various international projects such as: BALKANET – a multilingual semantic network of the Balkan Languages, EUROGLYPH – an Ideography-Based Code for European Communication, DigiCult-BG (project for digitization of national cultural heritage), and many others.

An important part of the Institute activities is dedicated to the compilation of various linguistic resources – large corpora of Bulgarian with original and translation text samples, Brown-designed and Tagged corpora of Bulgarian, etc., thus enabling comprehensive studies of linguistic phenomena. Another information source is the Institute's library to which generations of researchers have contributed for collecting its book-stock of more than 30000 books.

The services established at the Institute for Bulgarian Language contribute to the promotion of linguistic work among the public by implementing a range of services for other institutions, media, bureaus such as dictionary compilation, consultations, appraisals, etc. The Institute for Bulgarian Language founded awards in acknowledgement of the work and efforts of Bulgarian and foreign scientists.

The Institute for Bulgarian Language is publishing a number of magazines and series where the results of the work carried out at the Institute are presented. These are the magazines "Bulgarian Language", "Balkan Linguistics" as well as a number of series such as: "Proceedings of the Institute for Bulgarian Language" (since 1995 replaced by the series "Bulgarian linguistics"), "Work on Bulgarian dialectology and the history of the Bulgarian language", "Bulgarian lexicology and lexicography. Research and material."

Nowadays the Institute for Bulgarian Language is one of the most respectable academic institutions which has a great contribution to education and language competence.

---

### **Author Information**

Prof. Dr. **Vassil Rainov** – Director of IBL at BAS; Sofia 1113, 52 Shipchenski Prohod Blvd., bl. 17; Bulgaria  
e-mail: [rainovv@ibl.bas.bg](mailto:rainovv@ibl.bas.bg)

## COMPUTER PROCESSING OF MEDIEVAL SLAVIC SOURCES IN THE INSTITUTE OF LITERATURE AT BAS REPERTORIUM PROJECT (1994–2004)

Anissava Miltenova

---

### Introduction

Mixed-content miscellanies (very frequent in the Byzantine and mediaeval Slavic written heritage) are usually defined as collections of works with non-occupational, non-liturgical application, and texts in them are selected and arranged according to no identifiable principle. It is a "readable" type of miscellanies which were compiled mainly on the basis of the cognitive interests of compilers and readers. Just like the occupational ones, they also appeared to satisfy public needs but were intended for individual usage. My textological comparison had shown that mixed-content miscellanies often showed evidence of a stable content – some of them include the same constituent works in the same order, regardless that the manuscripts had no obvious genetic relationship. These correspondences were sufficiently numerous and distinctive that they could not be merely fortuitous, and the only sensible interpretation was that even when the operative organizational principle was not based on independently identifiable criteria, such as the church calendar, liturgical function, or thematic considerations, mixed-content miscellanies (or, at least, portions of their contents) nonetheless fell into types. In this respect, the apparent free selection and arrangement of texts in mixed-content miscellanies turns out to be illusory.

The problem was – as the corpus of manuscripts that I and my colleagues needed to examine grew – our ability to keep track of the structure of each one, and to identify structural correspondences among manuscripts within the corpus, diminished. So, at the end of 1993 I addressed a letter to Prof. David Birnbaum (University of Pittsburgh, PA) with a request to help me to solve the problem. He and my colleague Andrey Boyadzhiev (Sofia University) pointed out to me that computers are well suited to recording, processing, and analyzing large amounts of data, and to identifying patterns within the data, and their proposal was that we try to develop a computer system for description of manuscripts, for their analysis and of course, for searching the data. Our collaboration in this project is now ten years old, and our talk today presents an overview of that collaboration.

---

### 1994–1995

Bulgarian-American project "Computer Supported Processing of Old Slavic Manuscripts" begun in 1994, sponsored by IREX – Washington (1994–1995). A new type of software was built, which was based on the SGML (Standard Generalized Markup Language) accepted by the International Society of standardization (ISO) and especially in its TEI (Text Encoding Initiative) implementation. The goal of the project was to create a sophisticated system of processing Slavonic Manuscripts in the universal format with multiple using.

The system for computer analytical description of medieval Slavic manuscripts on the level of modern archeography, palaeography, codicology and textology (from now on – TSM = Template for Slavonic Manuscripts) was carried out in the process of the teamwork of David Birnbaum, Beirend van Dijk, who was then a post-graduate student in Groningen (The Netherlands) Milena Dobрева, Institute of Mathematics and Computing in BAS and Harry Gaylord, who taught computer systems in Groningen,. The experiments on the program, using tests, continued almost to the end of 1994. In July–August of 1995, the last changes and specifications in the system of document type definition (DTD) were made during the visit of David Birnbaum in Sofia together also with Andrey Bojadzhiev. The research project was sponsored also by the foundation 'Open Society', Sofia, in the period 1994–1997.

The description used here is specifically intended for the developing of a Repertory of the Old Bulgarian literature and letters and is adopted for Medieval Slavic texts. The development of fonts for writing the original texts in Medieval Cyrillic belongs to research associate Rumyan Lazov from the Institute of mathematics and computing in BAS. The searching programs on the second stage of the project were created by Stanimir Velev. The complex description of Slavic manuscripts is built by the standard of Standard Generalized Markup Language (SGML), which was accepted by the International Society of standardization (ISO). This electronic standard is based on the ability to include special "markings" in the texts of natural languages, so called tags. Tagging circles certain parts of the text and signal what the data represents. It makes very easy to draw out data from the text during its computer processing. This standard was used for the first time, in the description of Medieval Slavic manuscripts and for including an arbitrary (free from limitations) sizes of non-normalized texts from the manuscripts themselves in the

---

process of description. Our SGML-based undertaking was oriented not only toward preparing manuscript descriptions that might be suitable for printing, electronic rendering, and searching, as was the case with the database's approach. Rather, we anticipated even at that stage that the manuscript description files would be suitable for direct analysis, so that we would be able, for example, to identify patterns of structural similarity within a corpus of manuscripts on the basis of the same raw data files that we would also use to generate traditional printed manuscript descriptions.

The team has followed five main principles, formulated by David J. Birnbaum (see – <http://www.slavic.pitt.edu/~djb/>): 1. Standardizing of document file formats; 2. Multiple use (data should be separated from processing); 3. Portability of electronic texts (independence of local platforms); 4. Necessity of preservation of manuscripts in electronic form; 5. Orientation to the well-structured divisions of data according to the Slavic traditions of codicology, orthography, paleography, textology, etc.

The system for encoding of medieval Slavic text (TSM) was discussed on an international conference in Blagoevgrad (24th–28th July, 1995). The reports from the conference were published in a separate volume. The philosophy of SGML helped to settle some well known misunderstandings among palaeoslavists concerning philological questions of terminology, inventory of units, character sets and data structure.

---

### **1996–1999**

During the period from 1995 through 1998, a team of scholars supervised by me based primarily at the Institute of Literature at the Bulgarian Academy of Sciences produced SGML descriptions of some 200 medieval Slavic manuscripts of all types.

At the same time, the Institute of Literature entered into a project with Ralph Cleminson at the Central European University entitled "Computer-Supported Processing of Slavonic Manuscripts and Early Printed Books", which led to the encoding of additional manuscript descriptions and the publication of several articles addressing the technology underlying the project. Ralph Cleminson, David Birnbaum, and others presented the results of their research at the Twelfth International Congress of Slavists in Kraków in 1998, where the International Committee of Slavists established a Special Commission to the Executive Council of the Committee for the Computer-Supported Processing of Slavic Manuscripts and Early Printed Books, with David, Ralph, Andrey, and me as officers. The Commission's authorization was renewed at the Thirteenth International Congress of Slavists in Ljubljana in 2003. Participants from Belorussia, Bulgaria, Czech Republic, Finland, Italy, Macedonia, Great Britain, the US, etc. put on discussion some mainstream questions in the field.

The other principal achievement of this stage was the development by Stanimir Velev of a query interface for the manuscript descriptions that had been prepared within the Repertorium project. Stanimir's interface was an interim solution that has now been superseded by XSLT scripting, but for several years it served as the principal query engine for scholars at the Institute of Literature who were conducting philological research on the basis of our manuscript descriptions.

---

### **2000–2003**

For three years amount of analytically described manuscripts increased to three hundred. They were processed by using TSM system in the SGML environment with the corresponding interface A/E (Author/Editor, SoftQuad, Canada) software package. Members of the team were: Anna Stoykova, Nina Georgieva, Elena Tomova, Adelina Angusheva, Andrey Boyajiev, Margaret Dimitrova, Dimitrinka Dimitrova, Desislava Athanasova, Maya Petrova, Radoslava Stankova, Marina Jordanova, Dilyana Radoslavova, and Anissava Miltenova. The book under the title: "Medieval Slavic Manuscripts and SGML: Problems and Perspectives" (Sofia, 2000) is sponsored by IREX and Central European University). The articles in the book not only put into scientific circulation the achieved results from the analysis of the manuscripts, but also mark the problems that are waiting to be solved.

In general, the description in Repertorium is much more detailed in comparison with all other projects in the field, especially in such areas as orthography and the description of the texts in the manuscripts. The textological part includes information on the level of the whole manuscript (<manuscriptContentDesc> element) and on the level of each text (an element <articleContentDesc>). Because the intention was to provide research results from philological text investigations, there are elements such as <source>, <translation>, <protograph>, <antigraph>, and <litRedaction> on the level of the manuscript and on the level of each of the texts. A special element <neighbour> was introduced in the DTD to facilitate describing the organization of the component texts in mixed-content miscellanies, where the texts are not arranged according to ecclesiastical feasts or medieval typicons.

A current continuation of the original project, "Electronic Description and Edition of Slavic Sources" (2002–2003, sponsored by UNESCO), is in a transitional stage of migrating from SGML to XML technology. In 1994–1995, when the SGML DTD for the project was first constructed, Extensible Markup Language had not yet been conceived. Since then electronic and web technologies have changed very rapidly, and now we have tools that are very convenient for direct browsing and editing the markup files. Direct access to XML documents from such popular browsers as Internet Explorer, Opera, or the Gecko-engine powered ones, as Mozilla, Doczilla, and Netscape, provide more control and efficiency. This fact, together with the development of special recommendations for the markup languages produced under the auspices of the W3 consortium, Unicode, and other institutions and international initiatives, has led to a rapid growth of academic applications based on XML technology. So, this stage is characterized not only by the accumulation of still more manuscript descriptions, but also by the conversion of our materials from SGML to XML. The transition to XML was dictated by the remarkably broad acceptance of XML within the electronic-text community, and particularly by its adoption by the TEI, initially as an alternative to SGML, but ultimately as a replacement for it. We have currently converted over one hundred manuscript descriptions from our initial corpus of three hundred; the rest will be converted in time, and all new descriptions are being created directly in XML. Contributions of David Birnbaum to the project are enormous. His presentation at the Thirteenth International Congress of Slavists in Ljubljana, demonstrated a new level of document multipurposing: the generation directly from TEI XML manuscript descriptions of dendrograms that illustrated the degree of structural similarity among miscellany manuscripts and SVG plectograms showing the item-by-item correspondences in the contents of pairs of manuscripts. Andrey Boyadzhiev's presentation at the Ljubljana Congress illustrated the use of the same files to produce prose descriptions suitable for publication electronically or on paper.

At the beginning of Repertorium project we had concentrated on the production of manuscript descriptions based on the promise that one would be able to employ them directly in computer-assisted analysis at some point in the future. Last years had shown that the descriptions were suitable for use in a range of analytical applications, but primarily within a fairly low-level query framework that did not take full advantage of the hierarchical XML structure. David's work showed that radically new non-textual representations of manuscript structures were available essentially for free from the same files that were used to produce formatted descriptions. This development demonstrated that computers did more than provide a new way of performing such traditional tasks as producing manuscript descriptions. Rather, the production of electronic manuscript descriptions enabled new and innovative philological perspectives on the data. Not only did it make traditional activities easier and more reliable, but it also created opportunities for radically new philological research.

At the end of 2003 the project obtained a membership in Text Encoding Initiative consortium.

---

### **Ongoing Projects of the Department of Old Bulgarian Literature, Institute of Literature**

1. Joint projects with British Library, London, for analytical description of Slavic manuscripts (Working team: Anissava Miltenova, Andrey Boyadzhiev, Dilyana Radsolavova, Christine Thomas, Ralph Cleminson, with cooperation of Central Library of BAS – Dincho Kr"stev and Sabina Aneva).
2. Joint project with Gothenburg University, for implementation of computer tools in the study of late mediaeval Slavic manuscripts (working team from the Department of Slavic Languages and from the Institute of Literature BAS).
3. Collaboration with the Institute of Russian language, RAS, Moscow, directed to build a network for exchange of language and contents data of Slavic manuscripts in XML and to realize joint electronic editions.
4. The national project "Metadata and electronic catalogues" (Institute of Literature, Institute of Bulgarian language, Sofia university) is concentrated on the terminology in palaeoslavistic. The project includes something about 2000 denotations and keywords in the field of archeography, palaeography, textology, etc. which will be take into an electronic dictionary.
5. An ongoing project is an updating of my 1982 dissertation on the structure of mixed-content miscellanies. Our collaborative application of computer technology has already led us to revise some of my earlier conclusions about structural and textological similarities among manuscripts, and David Birnbaum and I are currently preparing a monograph that takes my dissertation as its starting point, but extends the data and the reevaluates it with the aid of computational tools.

---

### **Author Information**

**Anissava Miltenova** – Senior Researcher Dr., Institute of Bulgarian Literature, BAS, 1113 Sofia, Shipchenski prohod str. 52, Bulgaria; e-mail: [anmilten@bas.bg](mailto:anmilten@bas.bg)

## THE INVOLVEMENT OF INSTITUTE FOR INFORMATION TECHNOLOGIES IN TEXT PROCESSING

Georgi Gluhchev

*Abstract:* The activities of the Institute of Information Technologies in the area of automatic text processing are outlined. Major problems related to different steps of processing are pointed out together with the shortcomings of the existing solutions.

*Keywords:* Image Processing, Image Enhancement, Text Segmentation

### Introduction

The Institute of Information Technologies was created in 1994 as a successor of the former Institute of Technical Cybernetics, Institute of Engineering Cybernetics and Robotics and Institute of Informatics. Thus, it inherited the scientific traditions and investigations in modern areas like AI, Decision Support Systems, Multicriteria Analysis, Image Processing and Pattern Recognition, Information Processes, Systems and Media, Intelligent Systems and Soft Computing.

The problem of automatic text and handwriting processing is of great importance because its satisfactory solution will allow digitizing millions of printed and handwritten materials all over the world, thus making them broadly available. For example the problem for reliable and secure preservation of the cultural heritage is especially important and urgent one. This is why so many researchers all over the world have been involved in during the last decades.

Figure 1 represents the general scheme of an automated image processing system.

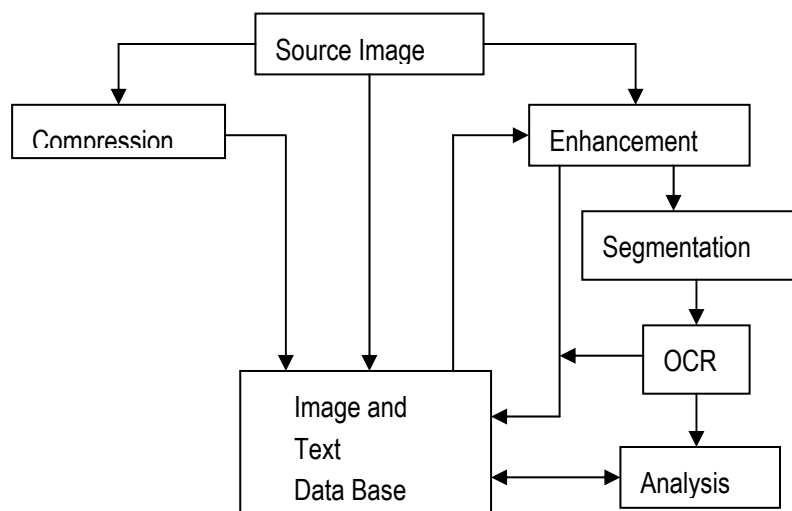


Fig. 1. General text image processing scheme

For more than two decades investigation work carried out by the Laboratory of Image Processing and Pattern Recognition has been aimed at the automatic processing and analysis of handwritten and printed text including letter recognition and writer identification. This led to the development and implementation of computerized systems for similarity estimation of handwritings mainly for forensic purposes. The Lab's research activities and application orientation are summarized in the following table.

Research areas	Application areas
1. Image processing	Handwriting analysis and writer recognition
Contrast improvement	Face recognition
Smoothed image restoration	Speaker recognition
Image segmentation	Moving objects tracking
Automatic and interactive feature extraction	Multimedia applications
Selection of optimal subset of features	Fast and reliable search in large data bases of images
2. Pattern Recognition	Recognition of car license plates
Linear classifiers	Printed characters recognition
Statistical decision rules	NN for robust control
Neural networks	Medical imaging
Clustering	

---

### Image Enhancement

---

Very often documents that have to be processed are of poor quality due to different factors. This holds especially for ancient manuscripts or printed texts where time and improper conditions (dry or humid air) or handling may cause severe damages, resulting in presence of random and structured noise and diminishing image contrast (Fig.2a). To make the image more pleasing visually, on the one hand, and suitable for further processing on the other, special image quality enhancement techniques have to be applied. Three groups of approaches aimed at noise reduction, contrast enhancement and line refinement could be outlined [Gluhchev G.][Pratt W. K.] . However it must be pointed out that improving image in one aspect may cause its deterioration in another. For example, noise reduction will diminish the image contrast and blur edges and vice-versa.

1. Noise reduction deals with different type of random or structured noise. To diminish the effect of random noise, variances of averaging or median-based filters are used.
2. Contrast improvement is aimed at the increase in color difference between the background and printed or written symbols. The corresponding theory includes methods based on dynamic range stretch, global or local histogram equalization with or without histogram clip [Pizer, S.M., E.P. Amburn, J.D. Austin]
3. Edge sharpening is aimed at the underlining of boundaries and strokes. The most popular methods are based on the evaluation of Laplacian or gradient in different directions. The unsharp masking is a computer variant of a well known photographic technique.
4. Line refinement is of great use when disruptions in symbol's strokes or stroke merge are present. In many cases significant improvement could be achieved if mathematical morphology operations such as erosion, dilation, opening, closing, skeletonization and gradient evaluation are applied.

---

### Image Segmentation

---

The goal of this operation is manifold [Shapiro, V., G. Gluhchev, V. Sgurev.].The first step is to extract text, i.e., to separate the text from the background. Provided the image is not too noisy and the background is uniform, a fixed threshold will be a fast and good solution. For such images the histogram is bimodal and the proper threshold corresponds to minimum between the two peaks. Unfortunately, in practice so nice images are rather exclusion than a rule. Very often images look like the one shown in Fig. 1a. In such cases the global threshold may either cause a significant loss of information or produce object-like artifacts, as shown in Fig. 2b. To avoid this, different locally adaptive methods have been developed, where specific threshold is evaluated taking into account the gray level distribution only in the predefined area, or in the neighborhood of every pixel. Fig. 1c demonstrates the effect of the application of the well-known Otsu method [Otsu N.].

The second step concerns the separation of rows. A very effective and cheap technique is based on the horizontal projections of the binarized image. The obtained shape has minimal values corresponding to the between-rows strips, while the peaks point out at the rows (Fig.3). To avoid false minimums due to ascenders or descenders, the shape has to be smoothed beforehand. The problem will aggravate if rows are not strictly

horizontal due to different reasons (Fig. 4). In that case techniques based on Hough transform [Lickforman-Sulem, L. and C. Faure] could be successfully applied. Thus, the angle of rotation could be evaluated and text rows might be rotated to become horizontal.

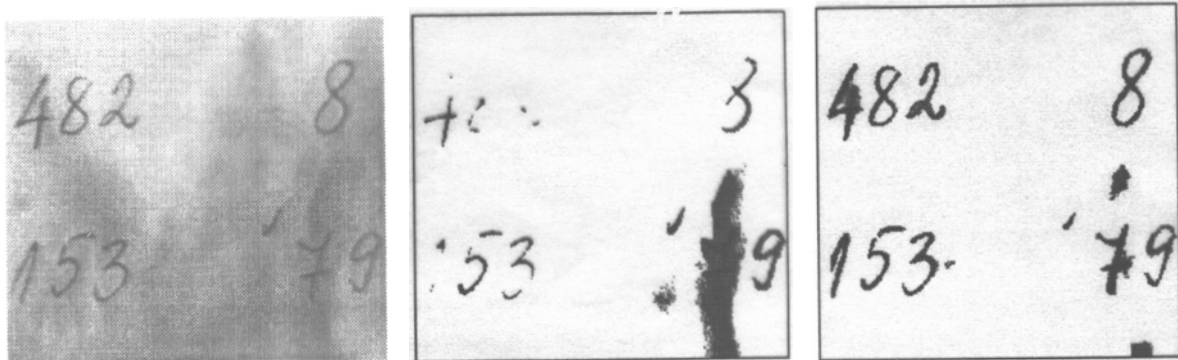


Fig. 2. a) Original noisy image                      b) Global threshold                      c) Locally adaptive threshold

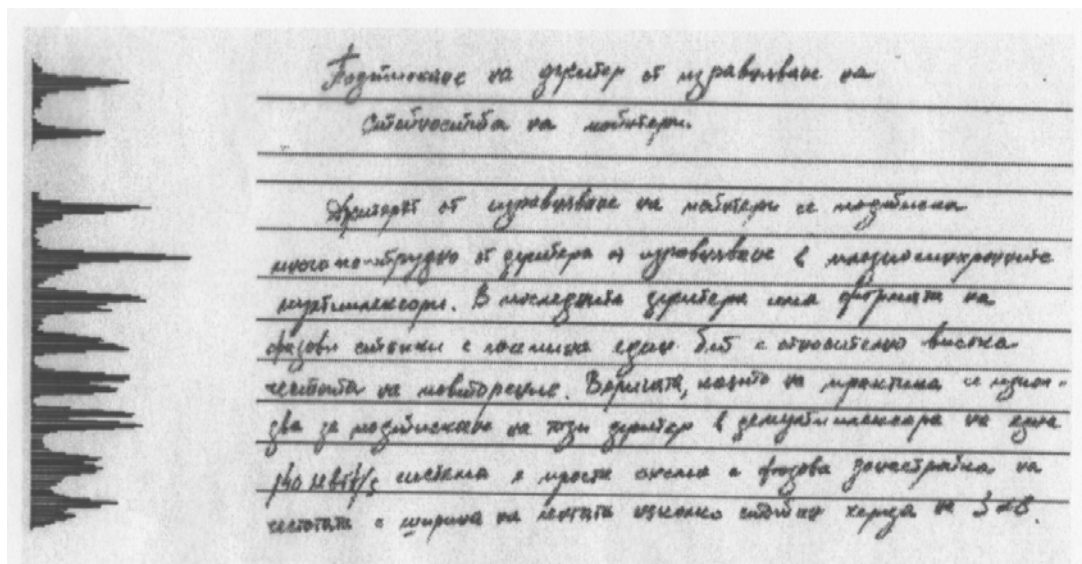


Fig. 3. Text lines separation

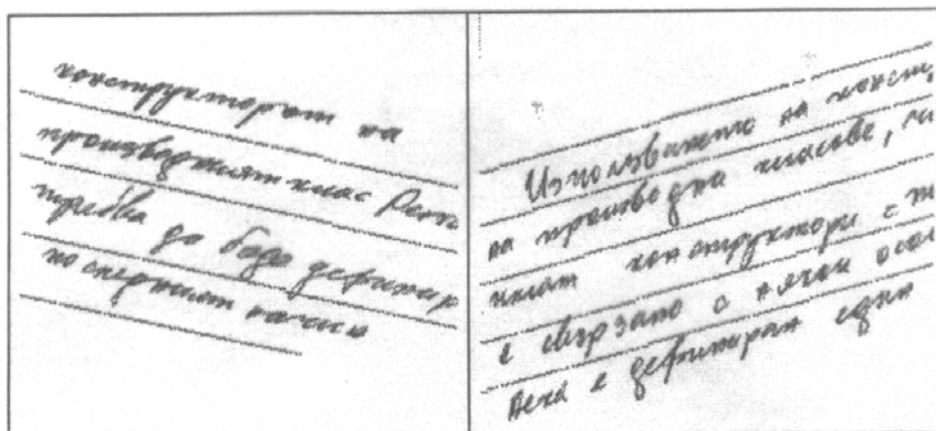


Fig. 4. Skewed text segmentation



The last segmentation stage concerns the separation of words, letters and strokes. While for the separation of words vertical projections of the rows could be successfully used, the problem of automatic segmentation of letters or strokes is quite complicated, especially for handwritten documents. Depending on the purpose of the investigation which may be text recognition, handwriting analysis or authentication, different processing is required. For example, in forensic investigations there is no need to recognize separate letters or words, but the goal is to establish document's or writer's authenticity. For this, specific features related to different letters, have to be measured and compared. They may include distances between specific points, angles, curvature, as shown in Fig. 5.

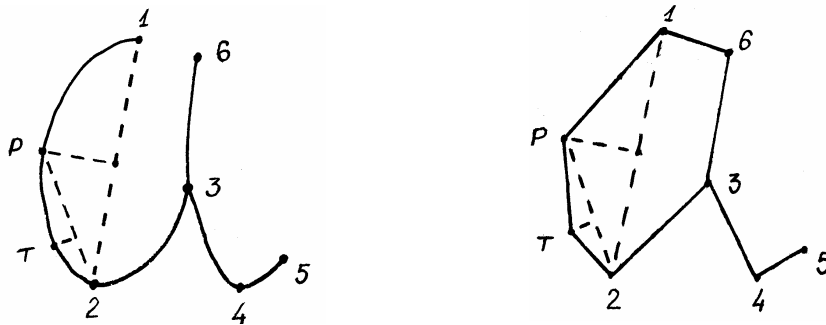


Fig. 5. Graphometric parameters of letter 'a'

Many approaches have been suggested for the recognition of words or letters. While the existing OCR packages perform very well on printed text, there is no reliable software for the recognition of scripts.

---

### Current Projects of IIT

At present the text's processing and analysis research work in the IIT is carried out within following projects.

1. BioSecure – Network of Excellence, from the 6th European Framework Program
2. Biometric parameters based identification – Contract No I-1302/2003 with the Ministry of Education and Science (MES)
3. Fast access methods by content to multimedia databases – Contract No I – 1306/2003 with MES
4. Method and software for effective search by graphical content in large data-bases of images – Contract No ID6/2003 with the Ministry of Transport and Communications.

---

### Acknowledgements

This work was supported by the Institute of Information Technologies and Ministry of Education and Sciences under contract No 1302/2003.

---

### Bibliography

- [Gluhchev G.] "Handwriting in Forensic Investigations", ICT&P, Varna, 2004 (in press)
- [Lickforman-Sulem, L. and C. Faure]. "A Hough Based Algorithm for Extracting Text Lines and Handwritten Documents", ICDAR'95, Montreal, 1995, pp. 774-777
- [Otsu N.] "A Thresholding Selection Method from Gray Level Histograms," IEEE Trans. Syst., Man, Cybern., vol.9, 1979, pp.62-66
- [Pizer, S.M., E.P. Amburn, J.D. Austin] et al. "Adaptive histogram equalization and its variations", Comput. Vision, Graphics and Image Proc., 39, 1987, 355-368
- [Pratt W. K.] Digital Image Processing, 2nd edn, John Wiley & Sons, 1991
- [Shapiro, V., G. Gluhchev, V. Sgurev.] "Handwritten document image segmentation and analysis", Pattern recognition letters, North Holland, 1993, 14, pp. 71-78

---

### Author Information

**Georgi Gluhchev** – Ph.D., Deputy Director of IIT, Acad. G. Bonchev Str., Bl. 2; Sofia 1113, Bulgaria.  
e-mail: [gluhchev@inf.bas.bg](mailto:gluhchev@inf.bas.bg)



---

## FACULTY OF MATHEMATICS AND INFORMATICS, SOFIA UNIVERSITY

**Maria Nisheva**

*Abstract: The Faculty of Mathematics and Informatics (FMI) of Sofia University "St. Kliment Ohridski" is briefly presented as an educational and research institution. The possible contribution of FMI to KT-DigiCULT-BG project is analyzed.*

*Keywords: computer science, information technologies..*

---

### Sofia University "St. Kliment Ohridski"

---

Sofia University "St. Kliment Ohridski" (<http://www.uni-sofia.bg>) is the oldest and most prestigious educational centre in Bulgaria offering over 80 BSc degree programmes and much more MSc degree programmes and PhD programmes in mathematical sciences, natural sciences and humanities.

Its academic staff consists of about 1800 lecturers including over 600 professors and associate professors. Sofia University has over 25000 students studying in 15 faculties.

Sofia University is one of the best scientific centres in the country. The most important characteristic of the research activities of Sofia University is their close relation with the educational process.

---

### Faculty of Mathematics and Informatics: General Information

---

The Faculty of Mathematics and Informatics (FMI) is an inheritor of the former Physico-Mathematical Department founded in 1889. In 1904 the Department was renamed as Physico-Mathematical Faculty and in 1963 it was separated as an independent faculty at the Sofia University. From 1986 it exists as a Faculty of Mathematics and Informatics. Now it is basically situated in the University campus "Lozenetz" (fig. 1).

FMI is responsible for the realization of teaching and research in the fields of Mathematics, Applied Mathematics, Computer Science and Information Technologies. It has over 2200 students and postgraduates.

The training process is provided by full-time lecturers (including about 70 professors and associate professors and more than 80 assistant professors) and by guest-lecturers who are usually well-known scientists from Bulgaria and abroad.

FMI has 14 departments and a lot of auxiliary sections (computer laboratories etc). The faculty staff and students have at their disposal a library with about 80000 volumes which contains the oldest collection of mathematical literature in the Balkan states.



Fig. 1: The main building of FMI

---

## Training Activities

---

### **BSc Degree Programmes**

FMI offers four BSc degree programmes that have been given the highest rating by the National Agency for Evaluation and Accreditation:

- Mathematics
- Applied Mathematics
- Informatics
- Mathematics and Informatics (provided for teacher qualification)

Three new BSc degree programmes in correspondence with ACM Computing Curricula 2001 are at different stages of preparation: Computer Science, Software Engineering, Information Systems. In particular, the Computer Science BSc degree programme started successfully in 2004/2005 academic year with a great number of extremely good candidates.

### **MSc Degree Programmes in Computer Science and Information Technologies**

FMI carries out the training in a significant number of MSc degree programmes. Most attractive are the programmes in the fields of Computer Science and Information Technologies:

- Artificial Intelligence
- Bioinformatics
- Computational Science and Engineering
- Distributed Systems and Mobile Technologies
- E-Business
- E-Learning
- Information Systems
- Logic and Algorithms
- Software Engineering

### **International Exchange of Students and Lecturers**

FMI participates actively in the programmes for international exchange of students and lecturers. Many staff members of the faculty realized useful long-term and short-term visits to famous European universities within the framework of a number of TEMPUS projects. Now students and lecturers from FMI use the opportunities for international exchange given by the SOCRATES/ERASMUS Programme. FMI has more than 20 bilateral agreements with universities in Germany, France, UK, Sweden, Norway, Denmark, Italy, Portugal and Greece for this purpose.

---

## Research Potential in Computer Science and Information Technologies

---

The academic staff of FMI doing teaching and research in the fields of Computer Science and Information Technologies includes:

- 19 associate professors
- 18 assistant professors (4 of them with PhD degree)

About 40 PhD students are also doing research in these fields.

The most effective areas of research in Computer Science and Information Technologies at FMI may be generalized in the following list:

- Programming Languages and Data Structures
- Markup Languages
- Databases and Information Systems
- Knowledge Based Systems
- Data Grids
- Virtual Reality
- E-Business
- E-Learning
- Mobile Technologies

---

## Project Activities in Computer Science and Information Technologies

---

FMI has rich experience in the management and implementation of research and development projects in the fields of Computer Science and Information Technologies. Some of the most successful projects of FMI in the last years are [Popov, 2003]:

- 5FP IST-1999-21148 (2000-2002) “Best Practice Pilot for the Implementation of Integrated Internet Based Remote Working Places for Virtual Teams Developing their Work at SMEs (IWOP)”
- 5FP IST-1999-20852 (2000-2002) “Best Practice Pilot for the Promotion and Implementation of Teleworking Tools at European SMEs of the Service Sector (PROTELEUSES)”
- 5FP IST-1999-12646 (2000-2002) “A Picture of Social Observation of Call Centre (TOSCA)”
- PHARE TEMPUS Institutional Building Project IB\_JEP-14047-1999: Centre of Excellence in Information Society Technologies
- 5FP IST-2001-34488 (2002-2004) EXPERT Project “Best Practice on E-project Development Methods”
- 5FP IST-2001-37460 COCONET “Context Aware Collaborative Environments for Next Generation Business Networks”
- 5 FP IST Project “DIOGENE: A Training Web Broker for ICT Professionals”
- 5 FP 2001/C 321/17 (2002-2005) GEM-Europe Project “Global Education in Manufacturing”
- 5 FP IST Project “WebLabs: New Representational Infrastructures for Learning”
- PHARE Multi-Country Programme in Distance Education (1998-1999) “DEMAND: DEsign, implementation and MANagement of telematics based Distance education”
- INCO-Copernicus 1445 Project “Flexible and Distance Learning through Telematics Networks: A Case Study of Teaching English and Communication and Information Technologies”
- INCO-Copernicus Project PL961125 under EC Directorate General XXIII “Intelligent Learning Environment for Course Telematics – INTELLECT”
- INCO-Copernicus Project 977074 LarFlast under EC Directorate General III “Learning Foreign Language Scientific Terminology”
- INCO-Copernicus Project 977102 ILPnet2 “Inductive Logic Programming Network of Excellence”

---

## Possible Contribution to KT-DigiCULT-BG Project

---

FMI may successfully contribute to KT-DigiCULT-BG Project with research and development activities in the following main directions:

- Development of software tools for creation, editing and visualization of catalogue descriptions of manuscripts and printed literature. Some promising results in this direction have already been achieved [Pavlov, 2004];
- Development of software tools for analysis of catalogue descriptions and digitized collections of written records using AI methods and techniques;
- Development of proper Web interface to electronic catalogues and digitized collections of cultural and scientific records;
- Providing a convenient access to the oldest collection of mathematical literature in the Balkan states.

---

## Bibliography

---

[Pavlov, 2004] P. Pavlov. XEDITMAN: A XML Editor for Manuscript Descriptions and its Implementation for Cataloguing of Bulgarian Manuscripts. In: Review of the National Digitisation Centre of Serbia and Montenegro, Belgrade (to appear).

[Popov, 2003] A. Popov (ed.). Research Activities of Sofia University “St. Kliment Ohridski” (2001-2003). Sofia University Publishing House, Sofia, 2003.

---

## Author Information

---

**Maria Nisheva-Pavlova** – Faculty of Mathematics and Informatics, Sofia University, 5 James Bourchier blvd., Sofia 1164, Bulgaria; e-mail: [marian@fmi.uni-sofia.bg](mailto:marian@fmi.uni-sofia.bg)

## TABLE OF CONTENTS

International Seminar "Digitization of Cultural and Scientific Heritage" .....	203
<i>Milena Dobрева and Nikola Ikonov</i>	
Digital Preservation and Access to Cultural and Scientific Heritage: Presentation of the KT-DigiCult-BG Project .....	205
<i>Mícheál Mac an Airchinnigh</i>	
The Experience at Trinity College Dublin .....	211
<i>Matthew Driscoll</i>	
The Experience of the Arnamagnæan Institute, Copenhagen .....	221
<i>Kiril Ribarov</i>	
The Latest Prague Contributions to Written Cultural Heritage Processing .....	224
<i>Stavros Perantonis, Basilis Gatos, Konstantinos Ntzios, Ioannis Pratikakis, Ioannis Vrettaros, Athanasios Drigas, Christos Emmanouilidis, Anastasios Kesidis, and Dimitrios Kalomirakis</i>	
Digitisation Processing and Recognition of Old Greek Manuscripts (the D-SCRIBE Project).....	232
<i>Giuliana De Francesco</i>	
MINERVA – the Ministerial Network for Valorising Activities in Digitisation Towards an Agreed European Platform for Digitisation of Cultural and Scientific Heritage.....	240
<i>Bernd Wegner</i>	
DML and RusDML – Virtual Library Initiatives for Covering All Mathematics Electronically .....	248
<i>Zdeněk Uhlíř</i>	
Manuscript Digitization and Electronic Processing of Manuscripts in the Czech National Library .....	257
<i>Yaşar Tonta</i>	
Integrated and Personalized Digital Information Services.....	263
<i>Boris Shishkov</i>	
Designing a Cultural Heritage Sector Broker Using SDBC .....	267
<i>Zoran Ognjanović and Žarco Mijajlovič</i>	
Digitization Projects Carried out by the Mathematical Institute Belgrade .....	275
<i>Charles Farrugia</i>	
Maltese Experience with Digitizing Cultural Heritage .....	278
<b><i>Representing of the Bulgarian Institutions</i></b>	
<i>Nikolay Markov</i>	
National Archives .....	282
<i>Elissaveta Moussakova and Alexandra Dipchikova</i>	
The Role of the National Library in Preserving National Written Heritage .....	284
<i>Vassil Rajnov</i>	
Institute for Bulgarian Language, BAS .....	288
<i>Anissava Miltenova</i>	
Computer Processing of Medieval Slavic Sources in the Institute of Literature at BAS Repertorium Project (1994–2004) .....	290
<i>Georgi Glushkov</i>	
The Involvement of Institute for Information Technologies in Text Processing .....	293
<i>Maria Nisheva</i>	
Faculty of Mathematics and Informatics, Sofia University .....	297