

SEVENTH FRAMEWORK PROGRAMME
FP 7-ICT-2007-1
ICT-1-4.1 Digital libraries and technology-enhanced learning

Large-scale integrating project

Annex I - “Description of Work” (public version)

Improving Access to Text
IMPACT

Project acronym: **IMPACT**
Project full title: **IMProving ACcess to Text**
Grant agreement no.: **215064**

This is the public abridged version of the Description of Work, Annex I to the Grant Agreement, of the IMPACT project. This version is intended for use outside of the consortium. Therefore some confidential information has been left out, but otherwise it is a copy of the original document. More information about the IMPACT project can be found at <http://www.impact-project.eu>.

FP7-2007-215064 IMPACT Description of Work - Public Version

<i>Part A Project Summary</i>	3
A 1. Project Summary	3
A 2. List of beneficiaries	4
<i>Part B Description of Work</i>	5
B 1. Concept and objectives, phasing, project architecture and list of tools and services	5
B 1.1 Concept	5
B 1.2 Objectives	6
B 1.3 Phasing.....	7
B 1.4 Overview of the project architecture	8
B 1.5 List of Tools and services	10
B2. Implementation	12
B 2.1 Consortium	12
B 2.2 Consortium as a whole.....	26
B 2.2.1 Other countries.....	27
B 2.2.2 Additional Beneficiaries	27
B 3. Expected impacts listed in the work programme.....	28
B 3.1 The significance of printed text	28
B 3.2 General challenges.....	29
B 3.3 Organisational challenges	30
B 3.4 IMPACT as a European Project	31
B 3.5 Centres of Competence.....	31
B 3.6 Horizons.....	32
B 3.7 Audiences	32

Part A Project Summary

A 1. Project Summary

Text that is not digital is virtually invisible. Today's readers search the internet for electronically accessible texts rather than visit the reading room of a library. Born-digital and digitised contemporary materials contain the richness that allows tools such as text mining and the semantic web to offer superior accessibility but the story is very different for historic documents. A vital part of the European heritage, encompassing more than four centuries of historic books and bound periodicals is becoming less and less visible to the public at large.

With the i2010 vision of a European Digital Library, the EU has launched an ambitious plan for large scale digitisation projects transforming Europe's printed heritage into digitally available resources. However, lack of institutional knowledge and expertise slows down the pace with which this vision can be realised. The state-of-the-art in OCR performance and machine understanding of the original document is inadequate, especially for historically important material with archaic fonts and spellings, newspapers with complex layouts, bound volumes, microfilm or typescript.

The IMPACT project will remove many of these barriers. It brings together fifteen national and regional libraries, research institutions and commercial suppliers - all centres of competence with unequalled experience of large-scale text digitisation processes and technologies. The project will let them share their know-how and best practices, develop innovative tools to enhance the capabilities of OCR engines and the accessibility of digitised text and lay down the foundations for the mass-digitisation programmes that will take place over the next decade. This project will facilitate a more collaborative approach to mass-digitisation. It will build capacity and lower the barriers to entry for organisations in the early stages of their own digitisation activity.

A 2. List of beneficiaries

Beneficiary Number	Beneficiary name	Beneficiary short name	Country
1 (Coord)	Koninklijke Bibliotheek	KB	NL
2	The British Library	BL	UK
3	Österreichische Nationalbibliothek	ONB	AT
4	Universität Innsbruck	UIBK	AT
5	Deutsche Nationalbibliothek	DNB	DE
6	Bayerische Staatsbibliothek	BSB	DE
7	Staats- und Universitätsbibliothek Göttingen	UGOE	DE
8	ABBYY Production LLC	ABY	RU
9	IBM Israel – Science and Technology Ltd	IBM	IL
10	Instituut voor Nederlandse Lexicologie	INL	NL
11	National Centre for Scientific Research "Demokritos"	NCSR	GR
12	Centrum für Informations- und Sprachverarbeitung, University of Munich	LMU	DE
13	University of Bath	UBAH	UK
14	University of Salford	USAL	UK
15	Bibliothèque Nationale de France	BNF	FR

Part B Description of Work

B 1. Concept and objectives, phasing, project architecture and list of tools and services

B 1.1 Concept

In the i2010 vision of a European Digital Library, the EU launched an ambitious plan for large scale digitisation projects transforming Europe's printed heritage into digitally available resources. The aim of fully integrating intellectual content into the modern information and communication technologies environment can only be achieved by full-text digitisation: transforming digital images of scanned books into electronic text. Over the last 2-3 years mass-digitisation has become one of the most prominent issues in the library world. Today, a number of advanced libraries in Europe are scanning millions of pages each year and large scale-digitisation is a matter of fact, not a vision any more. However, these efforts can tackle only a fraction of the total heritage available in cultural memory organisations. The digitised material is becoming available too slowly and in too small quantities from too few sources, for three reasons.

1. There is a lack of institutional knowledge and expertise which causes inefficiency and '*re-inventing the wheel*'. This is a problem for the vast majority of libraries, museums and archives in Europe.
2. The costs for full-featured electronic text of historical documents are much too high. Cultural heritage institutions will not be able to satisfy the needs of their users for electronic texts instead of pure digital images. Manual keying costs around 1 EUR per page, so that a typical book sums up to 400, 500 or even 1000 EUR.
3. Automated text recognition, carried out by Optical Character Recognition (OCR) engines does in many cases not produce satisfying results for historical documents. Recognition rates are poor or even useless. No commercial or other OCR engine is able to cope satisfactorily with the wide range of printed materials published between the start of the Gutenberg age in the 15th century and the start of the industrial production of books in the middle of the 19th century.

IMPACT as a network of centres of competence brings together fifteen national and regional libraries, research institutions and commercial suppliers. It will provide a wide range of highly innovative results in order to significantly improve this situation and lower the barrier for cultural heritage institutions to start mass-digitisation projects of printed material in a highly standardised, efficient and reliable way. It will do this by providing best practises, guidelines, help-desks, IT systems and demonstrators all dedicated to improving access to historical texts. Thus IMPACT will lay down the foundations for the mass-digitisation programmes that will take place over the next decade.

The vision that underpins the IMPACT consortium is that this is an extraordinarily complex challenge, spanning the entire workflow of a large-scale digitisation project: from initial decisions about digitisation strategy, image capture and pre-processing, full-text recognition using OCR, post-processing of results and enrichment of the document with linguistic, structural

and domain knowledge. The project will enable significantly wider public access to the material and help preserve the material for posterity.

B 1.2 Objectives

The main technical and research objective of IMPACT is, therefore, to drastically improve access to historical texts so that historical documents should wherever possible possess the same degree of accessibility as their born-digital counterparts.

In order to reach this ambitious goal IMPACT will focus on the following tasks:

- Develop OCR software and technologies which exceed the accurateness of current state-of-the-art software significantly, and which will allow for the first time to transform large amounts of digitised historical texts into electronic text.
- Provide a software system which will allow the realisation of new concepts of collaborative correction (in order to lower the costs for full featured full-text) by taking up and integrating the Web2.0 phenomena.
- Develop language tools and lexica in order to provide access to historical texts independently of historical variants of a given language.
- Support adopters of these tools so that more European historical lexica can be built.
- Develop a number of smaller modules such as image enhancement and segmentation toolkits, functional parsers, etc. in order to support the automated text recognition and/or access to historical text.

Apart from these clearly defined “hard” objectives IMPACT has also an important strategic objective: Support all European players such as libraries, cultural institutions, but also companies, decision making bodies, funding agencies etc. with high level information concerning the mass-digitisation and transformation of historical texts.

In order to reach this strategic goal IMPACT will focus on the following tasks:

- Build a network of competence centres in order to provide a single access point for all players involved in mass-digitisation and full-text generation.
- Set up guidelines and decision making tools.
- Organise workshops, training courses, and demonstrators for allowing other players to learn from the IMPACT results
- Provide a help desk and a knowledge base for resolving queries, channelling requests and collecting feedback.

To reach these goals, the IMPACT consortium will work for four years, following a work plan structured into four Sub-projects:

- SP1: Operational Context (SP-OC)
- SP2: Text Recognition (SP-TR)
- SP3: Enhancement & Enrichment (SP-EE) and
- SP4: Capacity Building (SP-CB).

B 1.3 Phasing

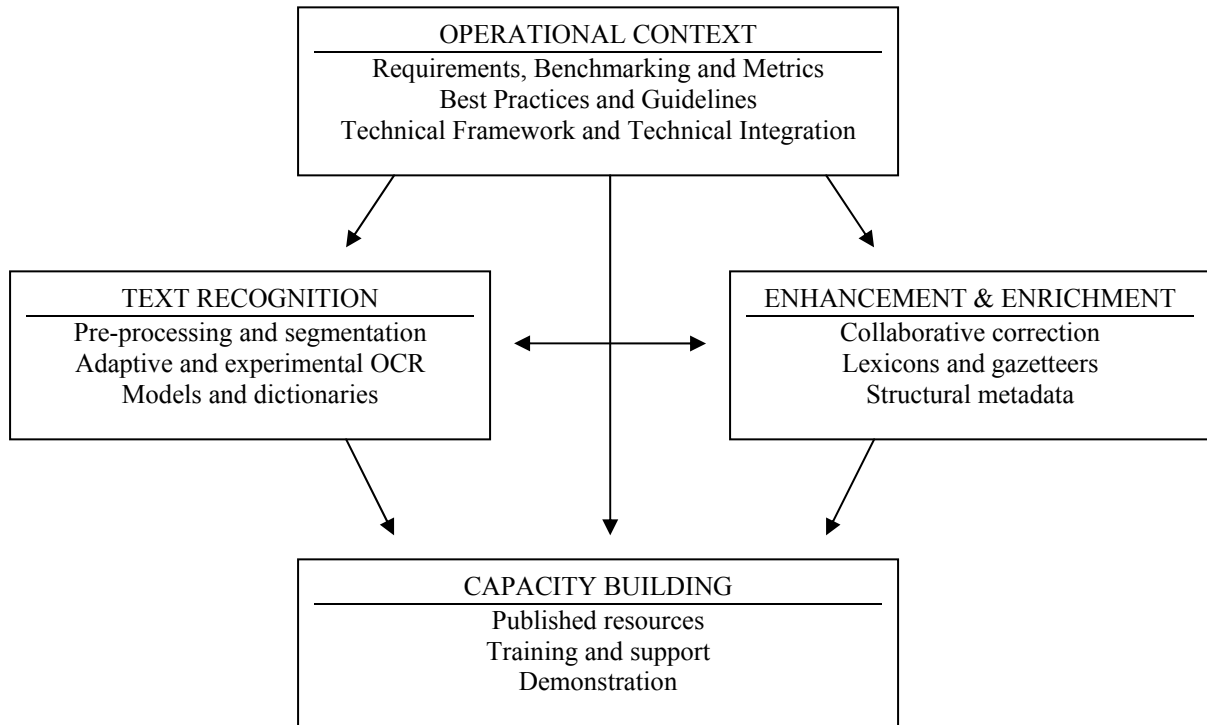
The project logically is divided into annual stage-plans, compatible with EC monitoring requirements. The first year contains an additional review-point, where the consortium can take stock of the initial activities across the project as a whole in month six and review the level of ambition in each work-stream.

The project plan rationale is summarised in the following table:

	Operational Context	Text Recognition	Enhancement & Enrichment	Capacity Building
Yr ½	Establish mechanisms for requirements definition and evaluation	Define benchmarks and establish research goal feasibility	Produce tool specifications and negotiate access to language resources	Establish communication channels and produce style-guide for project outputs Lay out a roadmap for the involvement of the second phase demonstrators
Yr 1	Publish workflow/cost models and guidelines. Define architecture for tool integration for demonstrators	Demonstrate experimental results for pre-processing, segmentation and, adaptive OCR	Demonstrate experimental results for all five Work Packages	First tools and guidelines published and Showcase demonstrator scope defined
Yr 2	Operational guidelines revised, technical guidelines produced, OCR performance comparison tool available	Evaluation of experimental techniques, integration of lexicon with OCR engines, OCR engines available for demonstrators	Early versions of all tools available for use in demonstrators	Technologies and support in place for demonstrators, web-hosted services implemented and Uptake demonstrator sites agreed
	Consortium work plan and budget reviewed, new participant(s) identified and proposal for change submitted to EC.			
Yr 3	Guidelines and case studies based on demonstrator evaluation	Tool enhancement programme concludes	Structural metadata and named entity work concludes, other languages supported through demonstrators	Support for Uptake demonstrators in place. Enhanced tools integrated and ready for use
Yr 4	Final guidelines, case studies and evaluation report	Final evaluation of results in the context of integrated adaptive OCR	Further work on additional languages	Support for second phase demonstrators

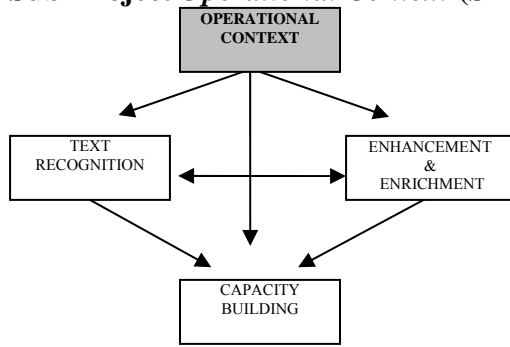
B 1.4 Overview of the project architecture

Removing the barriers for digitisation and improving access to digitised text presents many challenges. We will need an integrated approach that addresses research topics as well as the organisational requirements and the building of capacity in heritage institutions all over Europe. At the same time, a project of this size requires an overall project architecture with a clear structure that allows delegation of responsibilities. For these reasons, the Work Packages within the project have been grouped into four Sub-Projects which are defined below.



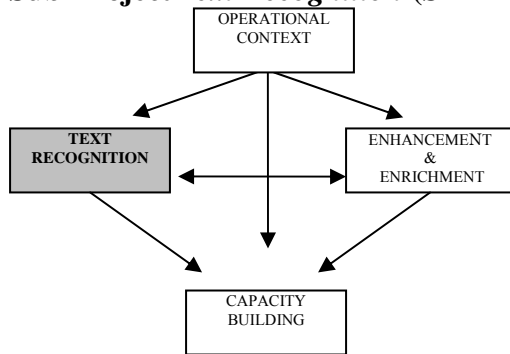
In addition, there are Work Packages dedicated to consortium and contract management, performance monitoring and coordination of dissemination.

Sub-Project *Operational Context* (SP-OC)



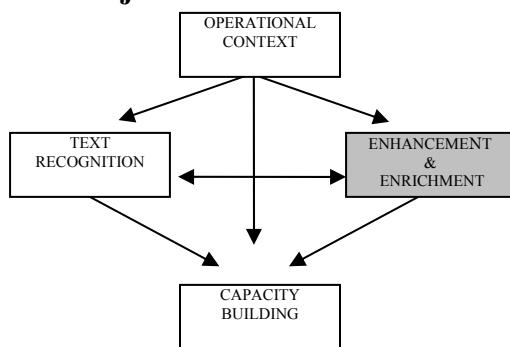
This Sub-Project will guide the direction of the project by representing the institutional view of mass-digitisation in terms of requirements, documentation of workflows and metrics for evaluation. It will also ensure the inter-operability of the discrete results from all parts of the project by defining an overall technical framework. It will integrate the tools and applications into a single managed collection of resources to enable end-users to interact with IMPACT tools and applications in a consistent manner.

Sub-Project *Text Recognition* (SP-TR)



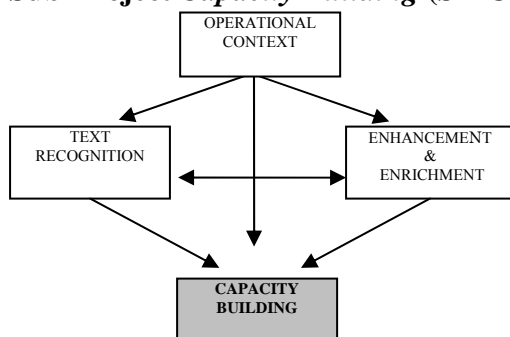
This Sub-Project brings together five research-oriented Work Packages concerned with the extraction of text in a digital form from the printed page. It will pursue short and longer term goals in improving OCR technologies, techniques for image enhancement and exploitation of linguistic knowledge.

Sub-Project *Enhancement & Enrichment* (SP-EE)



This Sub-Project is concerned with making the results of full-text digitisation more accurate and more accessible. It includes work on collaborative correction, descriptions of physical and logical structure of material and the use of historical and contemporary language lexica and named entities as ways of improving OCR performance and accessibility.

Sub-Project *Capacity Building* (SP-CB)



This Sub-Project stimulates and supports the uptake of results in the MLA community through a help desk and website and a dual-strand training initiative of work-placements and a vocational qualification. These results will be packaged to a consistently high standard suitable for publication to the target audience.

B 1.5 List of Tools and services

SP Operational Context:

1. Digitisation Decision and Planning toolkit: a set of models that can be used to initiate, organise, manage and cost mass digitisation projects. The models will be interactive and flexible to provide support in an intelligent way to the variety of institutions that are likely to make use of them.
2. Operational Guidelines: a set of documents which will provide guidance on real world implementation of the tools produced within the project.

SP Text Recognition:

Core systems

There are two core systems provided by the IMPACT project. Both are dealing with the OCR process itself. Whereas the ABBYY FineReader OCR engine represents the state-of-the-art in this field, the adaptive OCR system of IBM will be a cutting-edge contribution expanding today's technology significantly.

1. Adaptive OCR engine

A comprehensive software system which will improve the recognition of historical texts significantly by applying adaptivity as one of the main features to the text recognition process. It will integrate several other tools (see section on integration), such as the image enhancement toolkit, the ABBYY FineReader engine, the post-correction module and the lexica tools.

2. Improved FineReader OCR engine

The state-of-the-art OCR engine of ABBYY will be adapted in order to cope with the challenge of recognising historical fonts and layouts.

Minor systems, tools and prototypes

1. Image enhancement toolkit

A set of software tools for manipulating scanned images in order to improve the recognition results of OCR engines.

2. Segmentation toolkit

A set of software tools for recognising and segmenting important features of scanned documents, such as blocks, lines and characters.

3. Post-correction modules

A set of software tools for improving the lexicon based verification and correction of recognition results.

4. Experimental prototypes and tools

- A special OCR classifier capable of dealing with (especially hard to recognise) typewriter characters.
- A word spotting engine capable of searching words in texts which cannot be OCR processed in the traditional sense of the word.
- A research prototype for extracting the complete inventory (characters) of a given book based on shape clustering.

SP Enhancement & Enrichment :

1. Collaborative correction

A full web-based collaborative correction system including input management and user monitoring suitable for the wide public use. In addition a final report analyzing and summarising system operation for a number of niche applications will be provided.

2. Lexicon structure

A web service: Named entities repository and collaborative environment

The named entities repository is a collaborative web-based workspace for named entity management.

Tools

Toolbox for named entities variant resolution, matching and classification in historical documents

This toolbox provides the means to enable variant-independent named entity handling in the presence of historical language variation. The toolbox will contain:

- Tools for named entity classification
- Tools for named entity matching and variant resolution

Toolbox for building lexicon content

The toolbox provides the necessary computational tools to support the efficient lexicon building workflow. The toolbox will contain:

- Conversion tools for named entity data
- Tools for the extraction of named entity data from historical text
- Tools for construction of lexicon content from historical dictionaries and historical texts

Toolbox for lexicon deployment in enrichment

This toolbox will provide the means to overcome the historical language barrier imposed by spelling variation, historical inflectional and compounding principles by modern lemma assignment. The toolbox will contain:

- Tools for constructing and applying models of historical spelling variation
- A lemmatiser for historical text
- Tools for decompounding and dealing with non-standard word splitting practices

In addition, there will be a *toolbox of web-services for structural metadata encoding*

3. Lexicon building

- General lexica for Dutch, German and English
- Named Entities Lexica for Dutch, German and English
- Support for lexicon development in other European languages

4. Functional Extension Parser

With the Functional Extension Parser (FEP) users such as libraries, digitisation centers etc. will be able to exploit the standard output of OCR engines (XML) and to receive structural metadata such as page numbers, print space, default text size, and similar features. The toolbox will comprise a number of web-services with web-based GUIs for correcting the output.

SP Capacity Building:

1. An interactive project website that will provide access to all project outputs and form the nucleus of a virtual network of all European digitisation centres of competence and associated research activities
2. An established Help Desk system that will broker end-user requests to project partners and to other digitisation centres of competence
3. An established training programme dealing with large-scale digitisation issues and technologies, with a range of supporting documentation made available through the project website

B2. Implementation

B 2.1 Consortium

B 2.1.1 Koninklijke Bibliotheek (National Library) KB

The National Library of the Netherlands fosters the national infrastructure for scientific information and plays an important role in the permanent access to digital information at an international level. It has developed a vast experience in the digitisation of images and text since the mid-1990s and its e-Depot, the world's first digital archiving system for academic publications, now contains more than 7.5m articles. The KB also hosts the offices of The European Library (TEL), and of the European Digital Library (EDL). The work on IMPACT will be carried out within the Research and Development (R&D) Section of the KB which has taken part in many EC-funded projects over the years, e.g. TEL and the development of tools and services for digital preservation in the FP6 project PLANETS.

Expertise

Management of international projects; European infrastructure; Large scale text-digitisation of various formats. Research into enhancing results of OCR; development of models for workflow and quality control.

Goals

Enhance OCR for historical texts and 'difficult' fonts and layouts; to reduce cost by the use of better tools and streamlining the process of digitisation; providing meaningful access to historical sources. Speed up the tempo of digitisation in Europe by disseminating the results and so enlarging the corpus of the European Digital Library.

Main roles

The KB will coordinate the project and monitor the technical architecture and integration of the project.

Key personell

Team Leader: Hildelies Balk is head of the department of the National Programmes for digitisation in the R&D section of the KB. She holds a PhD in the History of Art and is an experienced researcher and project manager in the field of cultural heritage. Before joining the KB in 2006 she was head of collections in Museum de Fundatie. Hildelies coordinated the forming of the IMPACT consortium and the writing of the proposal.

Astrid Verheusen holds a Master's degree in history and has been working for the KB since 2001. She has many years of experience in research and development projects in the field of

digitisation and digital preservation, She is currently responsible for the digitisation of Dutch Newspapers (8 million pages) and in charge of several projects that support the technical infrastructure for all mass digitisation projects of the KB.

Marian Hellema has been working in the KB as a technical project manager since 2005. She is involved in all technical aspects of large-scale projects for digitisation. Marian developed the workflow, the metadata schema and the functional requirements that is used for digitisation in KB and is currently involved in the Dutch Digital Databank for Newspapers.

A technical project manager will be hired by KB for the work on technical architecture and integration

B 2.1.2 The British Library (National Library) BL

The British Library is the national library of the United Kingdom and one of the world's greatest libraries. The BL receives a copy of every publication produced in the UK and Ireland in a collection which includes 150 million items, in most known languages and grows by approximately three million new items every year. It has been involved in digitising items from its collection since 1993 and has developed great expertise in text and image digitisation of items created over the last 2000 years.

The Library has been involved as a partner in many projects funded at national, European and international levels. An example being The European Library (TEL), which started as a British Library led, European Commission funded project from 2000–2005, established with the express aim of giving the virtual user access to the digital and non-digital collections of the National Libraries of Europe. The Library continues to have representation on the Management Committee and steering group of TEL. Current projects include PLANETS (coordinated by the BL), EDL and UKWAC (web archiving)

The BL fully supports the European Commission strategy for i2010: the European Information Society and the development of the European Digital Library. The Chief Executive of the British Library is a member of the High Level Group chaired by EU Information Commissioner Viviane Reding and charged with developing the strategy around the European Digital Library.

The BL is committed to building a digital research environment a key element of which is the significant increase in the amount of digitised content available. To this end the Library has undertaken a number of mass digitisation projects which has made available three million pages of historical newspapers and books by 2009 the availability will grow to almost thirty million items

Expertise

Large scale digitisation of text-based materials including historical newspapers and books.

Utilising state of art for OCR and layout segmentation.

Workflow modelling, costing, producing guidelines and defining technical standards for end to end digitisation processes.

Goals

Advance the performance of tools for OCR and analysing layout of historical texts.

Reduce the per-page costs for mass digitisation of text based materials.

Develop tools which will allow the enhancement of extracted text and lead to a richer resource discovery experience for our users

Main roles

The BL leads the Sub-Project *Operational Context*

Key personell

Team Leader: Aly Conteh is currently working as the Digitisation Programme Manager at the British Library, a post he took up in April 2003. He is responsible for the development and implementation of policies to govern digitisation of items from the Library's collections in accordance with the British Library strategy. He currently acts as the Senior Responsibility Owner for the BL project that will digitise 25 million pages of 19th Century books from the BL's collections, in partnership with Microsoft.

Additional staff will be hired for the project.

B 2.1.3 Österreichische Nationalbibliothek (National Library) ONB

The Austrian National Library (ONB) is the main scientific library of the Republic of Austria. In addition to its role as a deposit library, ONB acts as a research centre and has been involved in numerous national and international digital library initiatives. The library has been a partner in several projects EC-funded in FP4, FP5 and FP6 as well as in the eContent and eContent*plus*. Recent projects include PLANETS, DELOS, BRICKS ,EDL and TEL*plus*. The Austrian National Library is involved in initiatives related to the i2010 Digital Libraries strategy of the European Commission. Since 2005 the library has been full member of The European Library (TEL). Within the eContent*plus* project EDL the Austrian National Library is leading the Work Package “*Developing The European Digital Library*” and is responsible for producing a digitisation roadmap for the European national libraries and for involving relevant stakeholders from the MLA community into the process of preparing the European Digital Library.

Expertise

EC Project Management experience, expertise in authority files. Digital object life cycle management. Contributor of rich historical material, large scale digitisation experience. Involvement in i2010 Digital Libraries initiative via TEL, EDL project, TEL*plus*, EDLnet and CENL Content WG.

Goals

Advance the performance and significantly reduce costs of OCR in large scale digitisation projects.

Develop tools to enhance text recognition rate and improve retrieval by employing computational historical lexica.

Main roles

The ONB leads the Sub-Project *Enhancement & Enrichment*

Key personell

Team Leader: Max Kaiser, MA, has worked for the Austrian National Library since 2000 and has been Coordinator for R&D Projects since 2004. He has many years of international experience in the field of digital library, digital preservation and digitisation research and leads the Testbed Sub-Project within PLANETS and acts as Work Package Leader in BRICKS. Within EDL he leads the “*Developing the European Digital Library*” Work Package. Max Kaiser is also coordinating the ONB's participation in The European Library (TEL).

Christa Müller, MA, studied at the University of Vienna where she participated in several international research projects. She has been working for the Austrian National Library since 1996 and has been Digitisation Programme Manager since 2002, responsible for planning, organising and controlling several mass digitisation projects (newspapers, journals, law gazettes and broadsheets). In 2006 she chaired a workgroup that developed the Austrian National Library's digitisation strategy for 2007–2011.

Joachim Korb, MA, joined the Research and Development team of the Austrian National Library in 2007. Currently he acts as a Work Package Leader of WP1 - "Making searchable digitised images via OCR" in the eContentplus project TELplus. Before joining the Austrian National Library he lead a digitisation team of the Central and Regional Library Berlin, Germany.

Georg Petz, MA, joined the Research and Development team of the Austrian National Library in 2007. He has a degree in informatics and is currently responsible for web service development and search engine optimisation in the EC project TELplus.

Bettina Kann - Bettina Kann, MA, has been working for the Austrian National Library since 1993. She has many years of experience in digital library projects. After working as a system librarian she has been head of the Austrian National Library's Digital Preservation department since 2004. She has been involved in the EC projects reUSE and PLANETS.

Additional staff will be hired for the project.

B 2.1.4. Universität Innsbruck (University & Library) UIBK

The University Library, Department for Digitisation and Digital Preservation (DEA) currently comprises 9 full-time equivalent staff members. Its main expertise is in project management and in the development of software applications in the fields of digital libraries and digitisation. The department has been responsible for the coordination of several R&D projects (all dealing with digital library issues) in the 4th and 5th framework programmes, as well as in the EU programmes eContent and eTEN. Central to DEA's role in IMPACT is the EU-funded METADATA ENGINE R&D project, which focused on OCR engines and structural metadata and contributes significantly to the baseline for IMPACT. Within WP-TR4 Experimental OCR Engines, we are joined by our colleagues from the Infmath Imaging Group in the Department of Computer Science at the University of Innsbruck. The group currently consists of 17 mathematical researchers whose work includes image segmentation/enhancing of 2D and 3D data, basic research in pattern and shape recognition and Thermoacoustic Tomography as a novel image acquisition technique. From a mathematical viewpoint these applications are strongly linked with techniques from variational calculus, partial differential equations (PDEs) as well as differential geometry which can be considered the group's main expertises.

Expertise

Management of international R&D projects for digitisation and digital preservation.

Image processing, mathematical modelling and image analysis.

Goals

Develop new and enhanced OCR engines for historical documents and design and develop a new OCR engine based on a radically new mathematical approach. Develop new and enhanced OCR engines for historical documents and design and develop a new OCR engine based on a radically new mathematical approach.

Main roles

Sub Project Leader for *SP Text Recognition*; development of Functional Extension Parser in *SP Enhancement and Enrichment*

Key personell

Team Leader: Günter Mühlberger, Ph.D., Head of the Department for Digitisation and Digital Preservation of the University Innsbruck Library. Professional experience: Coordinator and project manager of several R&D projects from the 4th and 5th EU Framework Programme. E.g. Project coordinator of the METADATA ENGINE project (2000-2003). Publications and lectures on digitisation issues. Project manager of national projects, e.g. Austrian Literature Online (one of the largest digital repositories in Austria) or Innsbrucker Zeitungsarchiv. Currently also acting

as coordinator of a multinational network with 14 partners from 9 European countries for a *Digitisation on Demand* service.

Senior technical advisor Albert Greinöckker, Mag. DI (Technical degree in computer science)

Current position: Senior Software developer at the Department for Digitisation and Digital

Preservation Professional experience: Technical coordinator of national and international software development projects. Focus on development of application software for digital repositories, digitisation workflows and webservices dealing with the automated extraction of text and image features.

Otmar Scherzer, Prof. Head of the Infmath Imaging group at the Department of Computer Science at the University Innsbruck. Infmath Role of the project: Modeling and prototype software for non standard application in character recognition and character inventory generation.

Experience: Local Coordinator EU Project DEIC, coordinator of a working group in the EU Project MACSI. Leader of several FWF Projects.

Markus Grasmair, Ph.D. Current Position: Research Assistant at the Institute for Computer Science Research experience: Researcher in different national projects in the field of applied mathematics with focus on variational calculus and mathematical imaging. Publications on imaging issues.

B 2.1.5 Deutsche Nationalbibliothek (National Library) DNB

The German National Library is the central archival library and national bibliographic centre for the Federal Republic of Germany. Its task, unique in Germany, is to collect, permanently archive, comprehensively document and record bibliographically without gap all German and German-language publications from 1913 on and to make them available to the public. The German National Library is the leading partner in developing and maintaining rules and standards in Germany and plays a significant role in the development of international standards. Three projects are particularly relevant to IMPACT:

Viaf: Matching and linking of the name authority files Personennamendatei (PND) and Library of Congress Name Authority File (LCNAF) to a VIAF (Virtual international Authority File).

MACS: Unmediated linkage between subject headings providing a subject search interface in German, English and French, transcending linguistic barriers in the three comprehensive subject heading authority files: the French Rameau, US Library of Congress Subject Headings (LCSH) and German Subject Heading Authority Files (SWD).

Crisscross: Extending the capabilities developed within MACS by linking the subject headings of the Subject Heading Authority Files (SWD) with the notations of the Dewey Decimal Classification (DDC) to provide a multilingual, user-friendly, thesaurus-based research vocabulary.

Expertise

Management of the German authority files.

Contributor of a variety of digitised and born-digital material. Research into automated text categorisation, in the fields of scanning and OCR-optimisation and in automated metadata extraction.

Expertise in consumer service and needs analysis.

Goals

Language independent retrieval for persons and names, for different language epochs.

Better recall and precision for digital material and building a basis for the semantic web.

Main roles

DNB leads the work on the helpdesk

Key personell

Team leader: Anke Meyer has a Masters degree in library and information science and works in the department for digital services at the German National library. She is responsible for the new online catalogue and service portal. Her key activities are the combination of the different services and catalogues under one single point of access.

Barbara Pfeifer has a diploma in library science. She works in the „Arbeitsstelle für Standardisierung“ (Office for Library Standards, AfS) at the German National library and leads the editorial work of the PND (Name Authority File). Her key activities are the coordination of the editorial board in the German speaking community and collaboration on development and standardisation for the PND.

B 2.1.6 Bayerische Staatsbibliothek(State Library) BSB

The Bayerische Staatsbibliothek (BSB; Bavarian State Library) is due to Germany's federal library system together with Deutsche Nationalbibliothek and Staatsbibliothek zu Berlin part of the "Virtual National Library". The department Digitale Bibliothek/Munich Digitisation Centre (MDZ) of BSB, established in 1997 with the financial support of the German Research Foundation, has since then built up significant expertise and capacity in the digitisation of different library materials from the 8th to the 21st centuries. So far, about 15 million files (28 Terabytes of data, long term stored at the Leibniz Computing Centre at Munich) have been produced as outcome of more than 60 own or cooperative projects. Continuing the VD 16 I project, MDZ will start in July 2007 a highly innovative mass digitisation project using scan robots (2007 EU-ICT-prize winner Treventus) for an automated scanning of 16th century books (VD 16 II), where about 40,000 volumes will be digitised in 24 months (approx. 7 million pages). The cooperation with Google Books for digitising material free of copyright (the start is scheduled for 2007) will effectuate an increase of the digital asset of a further 300 millions of pages. As MDZ aims at improving steadily the availability and the access to digitised material, OCR (either structured or plain text) and automated procedures of textual structure recognition is a main concern. Therefore MDZ is also participating in the projects "eBooks on Demand" and "MICHAEL PLUS", both funded by EU in the eTen-Programme.

Expertise

Large scale digitisation projects of text materials of various epochs.

Project realisation and management, development of work flow tools and presentation and retrieval software, long term preservation.

Goals

Pan-European level of text accessibility, improvement of mass digitisation and post-processing work flow for ancient print material.

Roles

BSB leads the work on the requirements forum

Key personell

Team leader: Dirk Scholz, M.A., is a staff member of MDZ since 2001. He has initiated and coordinated several national and international projects, being in particular responsible for the integration of new technologies. Since 2004 he is Head of Digitisation and Portals.

Margarete Wittke, Ph.D., staff member of MDZ since 1999, head of Production and Technical Infrastructure.

Fedor Bochow, M.A., staff member since 2006, is currently responsible for the implementation and work flow control of the EU-funded digitisation on demand project.

Additional staff will be hired for the project.

B 2.1.7 Staats- und Universitätsbibliothek Goettingen (University Library) UGOE

The Goettingen State and University Library (UGOE) is one of the largest libraries in Germany. On the basis of its background and tradition as a research library the UGOE was able to build up collections of national and international rank which were continually cared for and which scarcely suffered any loss or damage during the Second World War. It is this base from which its responsibilities as state library for Lower Saxony, special-subject collection library (e.g. Pure Mathematics) and the National Library for the 18th century stem. At the same time the UGOE is one of the leading institutions in the development of the Digital Library in Germany. In spring 1997 the Centre for Retrospective Digitisation in Göttingen (Göttinger Digitalisierungszentrum GDZ) was established to support the programme and coordinate national efforts towards standardisation in various fields (e.g. digital conversion, online access, bibliographic description). The Center in Göttingen is engaged in evaluation of tools and techniques for image capture and text conversion, bibliographic description, document management and the provision of remote access.

UGOE has 10 years of field experience in research and hands-on digitisation, is a founding member of several standardisation initiatives (e.g. METS), has experience in automated text categorisation with the *SWD* (German subject heading authority file) and in automated text encoding from raw OCR full-text to structured TEI (Project: Rise of Modern Constitutionalism). Since 1997, more than 5m pages spanning several centuries have been digitised by the GDZ.

Expertise

Large scale digitisation projects of various epochs (50% in gothic letters), development of models and tools for workflow modelling and quality control, service oriented architecture, presentation and retrieval, open standards, long term preservation.

Goals

Enhanced full text access to a variety of digitised documents,
Metrics for OCR quality assurance.
Framework of user-friendly, web-based services to the scientific community (outreach).

Main roles

UGOE leads the work on the project website

Key personell

Team leader: Ralf Stockmann has been head of the „Göttinger Digitisation Centre“ (GDZ) at the Göttingen State and University Library since January 2005. He has a degree in Social Science (Georg-August-University Göttingen) with main focus on media and communication studies. 1998 to 2001 he worked as Scientific assistant at the „Zentrum für Interdisziplinäre Medienwissenschaft“ (ZiM). Work field coordination, curriculum- development, public relations, course guidance service. Afterwards he coordinated the „Notebook-University“ project which emphasises knowledge management via Internet, he is also founder member of the „Stud.IP“ core-group (open source eLearning Software).

Team: Mahnke, Christian (TEI full text conversion, OCR orchestration)
Kothe, Jochen (User interfaces, content management systems, web services)
Enders, Markus (Full text document modelling)
Liebetruth, Martin (OCR quality assurance, metrics)

B 2.1.8 ABBYY Production LLC (Private) ABY

ABBYY is an international company with over 600 employees in 7 worldwide offices, and includes a Research & Development department of about 300 engineers and scientists in more than 15 groups located in ABBYY Headquarters in Moscow, Russia.

The ABBYY research area is artificial intelligence (AI) which includes linguistic direction and a full spectrum of document recognition, document conversion and data capture. ABBYY is an active member of key recognition scientific conferences such as ICDAR (International Conference of Document Analysis and Recognition), DAS Conference (Document Analysis Systems) and IWHRF (International Workshop on Frontiers in Handwriting Recognition).

ABBYY scientists periodically publish papers in that area.

ABBYY was part of a consortium of libraries and digitisation companies from across Europe who worked together in the METAe EC FP5 project which developed a unique OCR technology and the METAe Engine, a software package specifically designed to organise the workflow of the archiving and conversion of historical materials such as books, journals, magazines and newspapers.

Expertise

Research and Development expertise in OCR, document analysis , image enhancement and linguistic technologies.

Provides the technologies for various library and book scanning projects, including participation in FP5 Meta-e project, Soros funding librarian projects in Lithuania, Armenia and Russia.

Goals

Adaptation and enhancement of OCR, image pre-processing and document analysis technologies for better accuracy on historical documents.

Development tools for context and dictionaries adaptation of historical languages to raise the overall accuracy of OCR.

Main roles

ABBYY will provide the relevant OCR technology to enable research into improvements of image preprocessing, layout (or page) and analysis of documents. It will deliver OCR-XML output for the recreation of logical document structure, plus adaptive OCR and components for a full variety of different historical printing materials in European languages.

Key personell

Team leader: Konstantin Zuev, PhD (Head of Technology Group) has a vast experience in the fields of Document Analysis, OCR, semantic analysis, OCR and applied linguistics. He will be supported by top ABBYY project managers.

Dmitry Deryagin (Head of Document Processing Group) has a vast experience in document analysis and synthesis (recreation) of document layouts. He also has papers published in ICDAR. Vladimir Rybkin (Head of Character Recognition and Image Processing Group) has worked in the OCR area for 17 years and has made a significant input in evolving OCR technologies. He has also published papers on OCR.

Anton Masalovich (Head of OCR Subgroup) also works as a scientist in Moscow State University for OCR problems.

At some point in the future additional engineers might be added.

B 2.1.9 IBM Israel – Science and Technology Ltd (Private) IBM

IBM, the largest IT company in the world, includes a Research Division with 3000 employees in 8 labs around the world. The Haifa Research Lab (HRL) is the largest of the five labs outside the

United States. Since it first opened as the IBM Scientific Centre in 1972, the IBM Research Lab in Haifa has conducted decades of research that have been vital to IBM's success. R&D projects are being executed today by HRL for IBM labs in the USA, Canada, Europe and the Far East, in areas such as storage systems, verification technologies, multimedia, document processing, active management, information retrieval, programming environments, optimisation technologies, and life sciences. HRL staff members are actively involved in the academic community, publishing papers in leading conferences and journals, participating in programme committees, and organising conferences and workshops.

The IBM Haifa Research Lab is uniquely positioned to handle the research challenges presented in the area of document processing. The lab has proven contributions in the following related areas: automatic scan quality control, enhancement of scanned images, page layout analysis and segmentation (distinguish text from graphics), adaptive OCR (adapts to font and vocabulary of the document being analyzed), along with automatic and manual validation of OCR results.

Expertise

Research in the fields of: Image and document processing, enhancement of scanned images, page layout analysis, image binarisation, text segmentation, OCR for printed and hand written characters, logic-based correction, tools for efficient verification of OCR results..

Goals

Adaptive OCR system, that will automatically learn from its own errors in order to tune its resources to match the character of the specific document/book being processed.

Innovation in collaborative correction of OCR results.

Main roles

IBM leads the development of Adaptive OCR and Collaborative Correction

Key personell

Team leader: Eugene Walach PhD. - Chief architect, with over 20 years experience in different areas of image and signal processing, focusing on OCR, document processing, and error correction. Eugene has managed several large scale document processing projects. He holds over 20 patents, authored a book, and has published over 70 scientific papers.

Asaf Tzadok - Senior software engineer and expert in all aspects of image and document processing. Asaf was responsible for managing and developing an archive digitisation engine for the Hearst center in the USA.

Ami Ben-Horesh – Senior OCR engineer. Ami has over 30 years of experience in developing and consulting for OCR and document processing systems. He worked on the Israeli census where his contribution significantly improved the OCR results.

In addition, work will be performed by employees drawn from a larger pool of experts in the various specific fields that are needed in this project.

B 2.1.10 Instituut voor Nederlandse Lexicologie (National Research) INL

The Institute for Dutch Lexicology (INL) is a research institute financially supported by the governments of the Netherlands and (Flemish) Belgium. Its mission is to document the vocabulary and grammar of present-day and historical Dutch by creating, maintaining and improving the accessibility of language resources like dictionaries, corpora, computational lexica, thesauri.

There are approximately 50 employees working at the INL, divided over five departments. The work on IMPACT will be carried out by the Language Database Department and the EDP department. The Language Database Department has a long-standing experience in building corpora and lexica. Its major research project is the Integrated Language Database of the Dutch

FP7-2007-215064 IMPACT Description of Work - Public Version

Language from the 6th - 21st century, consisting of three major components containing corpora, dictionaries and computational lexica. The EDP department has extensive experience in the development of linguistic and lexicographical retrieval applications and linguistic processing and enrichment of language resources. The staff of EDP consists of 7 computer linguists, 3 software engineers and 3 system administrators.

INL has a central position in the management of Dutch language resources and technology. The INL has participated in the EC funded projects PAROLE, SIMPLE, TELRI, ENABLER and ELAN. The INL actively participates in the DAM-LR project on distributed access to language resources and is a member of the CLARIN common languages resource infrastructure network.

Expertise

Vocabulary and grammar of present-day and historical Dutch.

Creating, maintaining and improving the accessibility of language resources like dictionaries, corpora, computational lexica, thesauri; linguistic and lexicographical retrieval applications, linguistic processing and enrichment of language resources.

Goals

Expansion of corpus material suitable for linguistic research.

Efficient workflow and tools for creating and using lexical data in mass digitisation.

Main roles

INL leads the work on lexicon structure and lexicon content

Key personell

Team leader: Katrien Depuydt is head of the language database department. She is a historical linguist and lexicographer. She has worked on two major historical dictionaries and has many years of experience in managing electronic publishing and content management projects.

Michel Boekestein is a computational linguist. He holds a master's degree in language and speech technology and has extensive experience in managing NLP resources and in language modelling.

Bob Boelhouwer is a computational linguist. He holds a PhD: From letter strings to phonemes: the role of orthographic context in phonological recoding, 1998. Relevant experience e.g.: integration of lexical resources.

Jesse de Does is a computational linguist. He holds a PhD in applied mathematics and a master's degree in Slavic Linguistics, and has many years of experience in language processing and retrieval applications.

Tilly Ruitenbergh is a linguist and lexicographer. She has worked on corpora and lexica in major European projects.

Additional staff will be hired for the project.

B 2.1.11 National Centre for Scientific Research "Demokritos"(National Research) NCSR

NCSR is a self-governing research organisation, under the supervision of the Greek Government. It is the largest research organisation in Greece and is internally subdivided into 8 research institutes. The Institute of Informatics & Telecommunications (IIT) aims at playing a pivotal role in research and development (R&D), as well as in the transfer and exploitation of R&D results in relevant sectors of the Greek and international industry and society. The Informatics Department of IIT conducts research and develops infrastructure to support content filtering and extraction, adaptive intelligent systems, multimedia analysis and processing for cultural heritage applications including processing and recognition of historical documents.

Relevant previous experience includes National Projects (D-SCRIBE and POLYTIMO) on processing and recognition of historical documents and FP6 projects KT-DIGICULT-BG

FP7-2007-215064 IMPACT Description of Work - Public Version

(digitisation of cultural and scientific heritage), BOEMIE and SHARE on indexing and retrieval of multimedia information.

Expertise

Image processing and document image analysis, OCR, processing and recognition of historical documents, pattern recognition and multimedia processing, language and knowledge engineering.

Goals

Develop innovative methodologies in order to pre-process, segment and recognise document pages resulting from mass digitisation of archival machine-printed sources.

Main roles

In IMPACT, NCSR will coordinate WP-TR2 Segmentation and also will be involved in document image pre-processing (WP-TR1 Image Enhancement), experimental OCR (WP-TR4 Experimental OCR Engines) as well as on evaluation and quality assurance (WP-OC3 Evaluation Tools and Resources).

Key personell

Team leader: Basilios G. Gatos, PhD worked from 1998 to 2001 at Lambrakis Press Archives as Director of Research in the field of digital preservation of old newspapers. From 2001 to 2003 he worked in the field of document management and recognition at BSI S.A. as Managing Director of R&D Division. His main research interests at IIT are in image processing and document image analysis, OCR, processing and recognition of historical documents. He has more than 65 publications in journals and international conference proceedings.

Stavros J. Perantonis, PhD is Head of the Computational Intelligence Laboratory of NCSR and has authored more than 110 published papers. His current technical and research activities are in the areas of computational intelligence, pattern recognition and multimedia processing.

Ioannis Pratikakis, PhD has more than 50 publications in journals and international conference proceedings. His research interests include document image processing, 2D and 3D image analysis and content-based image / 3D models search and retrieval.

Anastasios Kesidis, PhD has more than 20 publications in journals and international conference proceedings. His research interests include image processing transformations, document image processing and medical image processing.

In addition, Ph.D. students, Georgios Vamvakas, Nikolaos Stamatopoulos, Georgios Louloudis and Konstantinos Ntirogiannis will work on the project.

B 2.1.12 Centrum für Informations- und Sprachverarbeitung, University of Munich (University) LMU

The CIS at LMU works on various areas of natural language processing, with many international contacts both in the academic and business world. It is well-known for its contributions to the fields of electronic lexicography, search engines, text correction, information retrieval and semantic search. The current focus of the group of Prof. Klaus U. Schulz is on adaptive techniques for improving OCR results and text correction systems, fast approximate search in dictionaries, as well as on the development and use of semantic knowledge bases for text enrichment and indexing. In all these areas, the group has published in many major international journals and conferences. Recently the CIS started a close collaboration with the historical language group of the University of Munich and with the Bavarian National Library in the field of historical documents and texts.

Expertise

Dictionaries and thesauri, fast approximate search in large dictionaries, adaptive text correction methods, semantic document indexing, document profiling.

Goals

Adaptive text correction methods for OCR on historical texts.

Special language resources (German) for OCR on historical texts and text enrichment.

Main roles

LMU leads the development of Language models and dictionaries

Key personell

Team leader: Prof. Schulz received his PhD in Mathematics at the University of Tübingen in 1987. In 1987-1988 he was a visiting professor at the University of Niteroi (Rio de Janeiro, Brasil). After his habilitation in Computer Science at the University of Tübingen he was appointed professor of computational linguistics at the University of Munich in 1991. Prof. Schulz was the leader of many national (DFG) projects and has participated in many international projects. He was the coordinator of the international EU project BILEDITA (mid 1990s). His main fields of interest are text correction, document processing, information retrieval and semantic knowledge bases. Prof. Schulz has more than 65 refereed publications in major international journals and conferences. In 2007, Prof. Schulz has the following memberships in programme committees of relevant international scientific events: IJCAI-Workshop on the Analysis of Noisy Unstructured Texts, International Conference on Information and Knowledge Management (CIKM), ACL Workshop Web as Corpus 2007.

Uli Reffle: Computational linguist, PhD student with an excellent background on finite-state technologies and their use for efficient and adaptive text correction. These topics will represent the kernel of the planned dissertation. Uli Reffle has already gained experience on various IMPACT-related topics in a project funded by the German Research Foundation (DFG) on adaptive text correction.

Annette Gotscharek: Computational linguist, PhD student with special background on the construction of electronic dictionaries and their use for text correction and text interpretation tasks. These topics will represent the kernel of the planned dissertation. Annette Gotscharek has already gained experience on various IMPACT-related topics in a project funded by the German Research Foundation (DFG) on adaptive text correction.

Additional staff will be hired for the project

B 2.1.13 University of Bath (University) UBAH

UBAH is based at the University of Bath, UK and is jointly funded by MLA: the Museums, Libraries and Archives Council and the Joint Information Systems Committee (JISC) of the funding bodies for higher and further education in England, Scotland, Wales, and Northern Ireland. Project funding is also received from the Engineering and Physical Sciences Research Council (EPSRC), JISC and the European Commission. UBAH also receives support from the University of Bath.

UBAH aims to inform practice and influence policy in the areas of: digital libraries, metadata and resource discovery, distributed library and information systems, bibliographic management and web technologies. It carries out research and development work, provides network information services, including the *Ariadne* web magazine, and runs a variety of workshops and conferences. UBAH carries out applied and technical research in key areas of interest to our core funders' stakeholder communities.

A number of themes underlie UBAH project work; namely research into the development and use of emerging metadata standards, open access to data and e-print repositories, digital preservation and the management of metadata schemas. These themes, central to the development of digital libraries, also support our core-funded work programme. UBAH is the UK Dublin Core Metadata

FP7-2007-215064 IMPACT Description of Work - Public Version

Initiative (DCMI) Affiliate Managing Agent and we also host the JISC W3C representative (UK Web Focus).

Expertise

Research into digital library development, repositories, digital curation, metadata, terminology services

Advisory role for UK cultural heritage and higher education organisations, part of the Digital Curation Centre.

Goals

Engagement with content holders and other stakeholders on digitisation best practice and state-of-the-art

Main roles

UBAH leads the Sub-Project *Capacity Building*

Key personell

Team leader: Dr Liz Lyon is the Director of UBAH, where she has been involved with the development and implementation of the common information environment. This includes building architectural models for distributed digital libraries and promoting synergies between digital libraries and Grid-enabled e-research environments. In this context, she has led the eBank UK project and is Associate Director (Outreach) of the UK Digital Curation Centre, in which UBAH is a partner. Dr Lyon also has a doctorate in cellular biochemistry.

Michael Day is Research Officer in the Research and Development team at UBAH. Since joining UBAH in 1996, he has worked on a series of externally funded research projects relating to metadata and resource description, semantic interoperability and digital preservation.

Additional staff will be hired for the project

B 2.1.14 University of Salford (University) USAL

The Pattern Recognition and Image Analysis (PRImA - www.primaresearch.org) Laboratory in the School of Computing, Science and Engineering at the University of Salford is an internationally distinguished centre specialising in research with real-world impact. PRImA was first founded at the University of Liverpool and in January 2005 moved to the University of Salford and expanded within the Informatics Research Institute (one of the highest ranked research institutes in the UK in Library and Information Management). For over 12 years, research has primarily focused in various aspects of Document Analysis and Recognition where innovations have earned PRImA significant academic standing and have found applications in Industry and other sectors. Projects ranging from the analysis and recognition of historical documents to the analysis of web documents have been funded by public bodies and Industry. More specific to the IMPACT project is the successful completion of the FP5 MEMORIAL (IST-2001-33441) project that produced a comprehensive toolkit environment for the digitisation and recognition of World-War II typewritten documents. PRImA led the development of the Document Image Analysis tools. Research actively continues in the challenging field of the analysis and recognition of degraded historical documents. In addition, PRImA has developed (and continues to work on) methods and datasets for in-depth evaluation of Layout Analysis methods. Since 2001 PRImA has jointly organised (with the National Centre for Scientific Research of Greece – NCSR) the first and longest-standing series of international competitions in Layout Analysis.

Expertise

Research and Development in Image Analysis and Pattern Recognition with particular emphasis on Document Image Analysis, Recognition and Performance Evaluation. Significant experience in the analysis of historical documents.

Goals

Development of improved methods for Image Enhancement, Segmentation and typewritten OCR. Development of large-scale test dataset and improved performance evaluation techniques for Document Analysis and Recognition.

Collection of case-study material for teaching in Document Analysis and Engineering.

Main roles

Usal leads the work on benchmark definitions and image enhancement

Key personell

Team leader: Dr Apostolos Antonacopoulos is a Senior Lecturer and Director of PRImA. He received his PhD from the University of Manchester Institute of Science and Technology (UMIST), UK in 1995., He has worked and published extensively on various problems in Document Analysis and in Pattern Recognition and applications. He is a member of the Editorial Boards of the International Journal on Document Analysis and Recognition (IJ DAR) and the Electronic Letters on Computer Vision and Image Analysis (ELCVIA) journal, and Chair or member of a number of high-profile International Committees. He is a member of programme committees of most conferences in the field of Document Analysis and Recognition and he has recently (May 2007) co-edited the first special issue on the Analysis of Historical Documents in IJ DAR. He has significant experience in leading and participating in national, European and industry-sponsored projects. He was responsible for the document image analysis work on the MEMORIAL (IST-2001-33441) project.

Additional staff will be hired for the project.

B 2.1.15 Bibliothèque nationale de France (National Library) BnF

The Bibliothèque nationale de France (BnF) is one of the largest public and research libraries in the world. Its digital library *Gallica* (<http://gallica.bnf.fr>) contains 86,000 printed and 250,000 iconographic materials, obtained through the library's commitment to the digitisation of selected items of its collections. The current 19th Century Newspaper Digitisation Project is bringing more than 3 million pages available both in image and text mode. Moreover, 100,000 materials per year during 3 years will be added. In preparation for the new version of *Gallica*, planned for late 2007 with new and modern functionalities drawing upon the most recent Web 2.0 experience, more than 60,000 documents (out of the 86,000 presently available) will be converted to text mode through OCR software. The experience gained through *Gallica* has led the BnF to develop *Europeana*, a mock-up, followed by a prototype of a European digital library, which has been put online on March 2007. Although developed in only five months, this prototype has provided interesting experience on full-text indexing and customised services as well as cooperative work with the national libraries of Hungary and Portugal. The BnF coordinates the International Internet Preservation Consortium which aims at sharing experiments and developments for selecting, harvesting, collecting and preserving as well as providing access to internet content now and in the future. In 2007, 25 national libraries as well as the American foundation Internet Archive are involved in this programme.

The BnF is a founding member of The European Library consortium. The BnF is also involved in the *TELplus* project, in which it will explore the high quality OCR, full-text indexing and subject multilingual issues, as well as in the network of excellence *EDLnet*.

Expertise

Experience in the fields of digitisation, accessibility to and preservation of digital resources as well as in authority files. Gallica digital library.

Development of the prototype Europeana which will contribute to the European Digital Library. Participation in METAe project, Quaero project, TELplus, EDLnet as well as the International Internet Preservation Consortium.

Goals

Tools to enhance mass OCR for 17th-18th century printed materials in order to improve better access.

Roles

BNF is involved in several Work Packages in *Operational Context* and provides demonstrators

Key personell

Isabelle Dussert-Carbone, Head of the Preservation and Conservation Department since September 2006. This department is in particular in charge of digital preservation and mass digitisation programmes besides the IT and Digital Library Departments.

Marie-Elise Fréon, Head of the Digitisation Unit, Preservation and Conservation Department.

Laurent Duplouy, Production Manager for Digitisation and Preservation, IT Department.

B 2.2 Consortium as a whole

The IMPACT consortium brings together a critical mass of memory institutions with experience from previous collaborations, unique collections that provide specific challenges and staff with particular expertise relevant to the project. The collection-holding partners within the consortium have strong competence in the digitisation of text based materials. They also have experience of servicing requests to share that knowledge with other institutions who seek to develop their own competences. They share the vision of the contribution of a European Digital Library in shaping the cultural and social fabric of Europe for the foreseeable future. All of the memory institutions in the consortium have previously collaborated on TEL/EDL and BL and DNB are represented on the High Level Experts Group. Very recently BnF developed Europeana, a mock-up, followed by a prototype of a European digital library, which has been put online on March 2007. This core of user organisations will be expanded during the demonstrator phase of the project (see 2.3.3.1 below) by a number of other collection-holders.

There are also eight research institutions in the consortium, selected not only for the relevance of their experience and interests but also for their ability to play multiple roles within the project and thus reduce the complexity of the consortium from a management perspective. UIBK for instance, who brings considerable skills in image processing, mathematical and image analysis as well as experience in the management of international project for R&D to the project, not only leads the Sub-Project *Text Recognition* and develops tools in the Sub-Project *Enhancement & Enrichment*, but is also involved in the technical framework and technical integration in the other two Sub-Projects, thereby providing indispensable links throughout the project. Two of the research partners have a strong computational linguistics group. Both partners have enough technical expertise to be able to interface the work done with the work in text recognition. LMU has ample experience in using lexica and various finite-state techniques to improve text recognition and in building morphological lexica, also for historical text. INL has ample experience in historical linguistics and management of historical corpora and lexica and both technical and linguistic aspects of lemmatisation.

Six of these research teams are from the public sector and will work alongside two private sector companies with a strong reputation for research and a market-leading presence in document digitisation. The commitment of our private sector partners, ABBYY and IBM, establishes a sustainable future for the tools that the project creates.

All partners share the recognition that the private sector has an important role to play in that vision. This is true of large corporations (BSB are working with Google, BL with Microsoft, BnF with France Telecom regarding the organisation of documents, the search and results functionalities and the treatment of natural language applied to texts) but it is also relevant to consider the market opportunities for small innovative start-up companies such as the Austrian university spin-off company, Treventus Mechatronics. They won the European ICT Prize at CeBIT after just a year of trading for their page-turning product which is in use at UIBK and BSB. Such small companies are at the opposite end of the SME scale to our partner ABBYY (with over 600 employees, half of whom are deployed in R&D) which has the capacity to play a full role in this large 4-year research project. Support for SMEs can be found in the emphasis in the *Capacity Building* Sub-Project that is being placed on technology-transfer and support for innovation.

B 2.2.1 Other countries

We are pleased to have IBM (Haifa) from Israel (an Associated Country) and ABBY from Russia (an International Cooperation Partner Country) in the consortium but there are no participants from ineligible territories.

B 2.2.2 Additional Beneficiaries

It is intended that there should be an expansion of the consortium prior to the major demonstrator phase of the project. In month 6 a roadmap for selecting and involving the second phase demonstrators will be laid out.

The methodology for ensuring a cost-effective, transparent and objective way of expanding the consortium will be agreed with the Commission during year two.

B 3. Expected impacts listed in the work programme

The main level of ambition of the IMPACT project is to give digitised text-based documents a functionality that is equivalent to their born-digital counterparts.

The project does not address related research activities in areas such as digital preservation or innovative access services based on semantic search but its results will be significant enablers for current and future work in those areas. As an integrating project, it may well provide a suitable platform over the medium term for work in areas such as automatically summarising and/or assigning content-based metadata or keywords to documents.

According to the FP 7 Work Programme, IMPACT will contribute to two main objectives of ICT-2007.4.1 Digital Libraries and technology-enhanced learning. These are:

- stimulating the migration of content to digital form by improving the cost-effectiveness of OCR processing of printed material and the digitisation process as a whole, and
- improving the accessibility of that content once it is digitised and incorporated within the European Digital Library through enrichment of the digitised text, the addition of linguistic annotation, and techniques for describing the logical and physical structure of documents.

B 3.1 The significance of printed text

Printed text represents the intellectual heritage of Europe. Today access to this heritage means providing users with the chance to search, copy, paste, tag the full-text of a historical text. Although many Member States have digitisation programmes, efforts are fragmented and progress has been relatively slow. For example, the situation at the British Library is that its collection of some 150 million items is growing by 2% per year, but over the last ten years only around 1% of the collection has been converted to full-text information¹.

The Council Conclusions on the Digitisation and Online Accessibility of Cultural Material, and Digital Preservation (2006/C 297/01) underlines that:

“while making from the outset conceptual and technical preparations for all categories of cultural material (texts, audiovisual, museum objects, archival records etc.), the European Digital Library may exploit in its early stages the potential of a critical mass of multilingual textual material”

The concept of critical mass is important. The existing service The European Library (TEL) is an ambitious pioneering collaboration between European national libraries supported by the EU and created under the auspices of CENL, the Conference of European National Librarians. The European Library’s importance is specifically recognised and mentioned in COM(2005) 465 final, the Communication on digital libraries. TEL offers a professionally designed and maintained single access point and by the time IMPACT is under way early in 2008 TEL will give access to the digitised collections of all EU national libraries amounting to some 2 million digitised items.

¹ Approximately 3.5 million pages have been successfully OCRed

The i2010 Digital Libraries vision is to treble the volume of material online between 2008 and 2010. CENL recognises that “*the sheer existence of a massive corpus will trigger more digitisation as well as technical research and development*“. Critical mass is also likely to be more attractive to private partners and the project will help “*to kick-start the process of digitisation*” in smaller memory institutions at sub-national level and facilitate a more collaborative approach to full-text digitisation.

The project will lower the barriers for organisations to participate in the early stages of their digitisation activity and facilitate continual improvement in full-text digitisation activity within the community. A network of centres of competence in full-text digitisation will provide a framework which institutions can use in undertaking their digitisation activities and provide research solutions to issues which inhibit high productivity in digitisation projects. As already indicated, the basis for this will be laid within Sub-Projects *Operational Context* and *Capacity Building* where the MLA community will pro-actively be tackled and provided with highly valuable information material, information services and last but not least software tools and systems.

Essentially, the project is a measure designed to speed up the digitisation lifecycle, taking consideration of the availability and allocation of funding as well as the time taken to create the resources. It will provide ways of improving performance especially for challenging material such as newspapers with complex layouts, historical volumes with archaic fonts and language and poor quality microfilm and scanned copies of material. It will identify and deliver documented approaches to digitisation of such material which define best practices; clear, streamlined processes capable of delivering the material types and type of resources that users and private partners require, and relevant automated tools to support those processes.

B 3.2 General challenges

Full-text digitisation is labour-intensive and costly. It takes a considerable upfront investment, which in most cases goes beyond the means of the institutions holding the content. Offshore re-keying of important texts will remain an option² until OCR can deliver the required level quality for material that is currently difficult to process (i.e. some 60-80% of all public domain material published between 1500 and 1900 and hosted by libraries).

However, as recognised in Commission Recommendation (2006/585/EC): “Investments in new technologies and large scale digitisation facilities can bring down costs of digitisation while maintaining or improving quality and should therefore be recommended.” Efficient image pre-processing, high-performance OCR and collaborative correction are three of the areas where investment in quality-improvement can also deliver cost-reduction.

The results of the project will:

- Improve price/performance (especially for difficult source material and in institutions with very limited expertise) and make the business case for full-text digitisation more attractive. This will be realised mainly in the Sub-Projects *Text Recognition* and *Enhancement & Enrichment* where the following deliverables will contribute to this objective:
 - Image enhancement toolkit will allow to improve scanned images in order to support OCR engines to process “difficult” historical material
 - Segmentation toolkit will allow the processing of historical documents with sophisticated layout, such as newspapers or popular journals

² Costs are currently around 60–80 EuroCent for 1000 characters; a typical book page has 2000 characters, so around 1,2–2 EUR per page for structured full-text typed in China or India

FP7-2007-215064 IMPACT Description of Work - Public Version

- Adaptive OCR will significantly improve the processing of historical texts through the exploitation of self-learning technology and the improvement of state-of-the-art OCR engines (ABBYY FineReader)
- Improve access to historical texts by overcoming language barriers. Due to the fact that text recognition relies heavily on language processing the IMPACT project emphasises this aspect and provides a number of tools specifically designed to cope with this challenge:
 - Post correction system will improve the language processing in OCR engines and therefore significantly contribute to a enrich the output of text recognition.
 - Deliverables dealing with the creation and management of historical lexica and named entities which will be the bases for many new or improved digital library services, such as full-text search across different domains and centuries.
- Provide new options for correcting results coming from OCR engines. Since correction is one of the most labour-intensive and therefore costly processes within the generation of full-text the IMPACT project sets up innovative methods and tools to overcome this bottleneck.
 - Collaborative Correction is one of the main attempts to provide new tools and interfaces for correction as well as to involve the user-community.
- Provide access to the expertise of the project partners as centres of competence and encourage Member States to set-up and sustain large scale digitisation facilities.
 - Especially work carried out in Sub-Project Operational Context will be developed in order to support the MLA community with valuable tools to manage their digitisation efforts.
 - In addition in Sub-Project Operational Context will provide more specific advice for the use of modules and tools developed by the IMPACT project. “Knowledge base” will be one of the helpful outputs which addresses the needs of the MLA community.

B 3.3 Organisational challenges

New ways of working are necessary to make digitisation happen. Investments in digitisation must be accompanied by organisational changes within the institutions concerned, including upgrading the skills of the staff involved. One study³ across several digitisation programmes from the 1990s suggests that actual digitisation costs and creation of metadata amount for just 61% of the cost of a digitisation programme, the rest being organisational issues such as quality control, selection of material and rights clearance. In the case of coordinating partner, KB, organisational issues consume the majority of the budget of their current text-digitisation programme.

The results of the project will:

- Deliver a coherent programme of dissemination, training and demonstration aimed at capacity-building in and beyond participating institutions.
- Provide Best Practice guidance about the operational context for digitisation and removing constraints to digitisation. It is essential that material to be digitised is selected in order to present thematic and coherent collections rather than simply material with which it is easiest to meet digitisation targets.
- Our work in this area is complementary to the agenda set by the National Representatives Group (NRG) to develop mechanisms to promote good practice and skills development in accordance with the Dynamic Action Plan.

³ Puglia, Steven. "The Costs of Digital Imaging Projects" - *RLG Diginews*, (Vol. 3, No. 5)

B 3.4 IMPACT as a European Project

Developing a European Digital Library is by definition a European undertaking. It could not be done at national level at all both because of its nature and because the European partnership involved allows the necessary financial and human resources to be brought together. Research institutes from the United Kingdom, the Netherlands, Germany, Austria, France, Greece, Israel and Russia work together in this project as the collective expertise to tackle the problems to be addressed cannot be found within a single Member State.

The initial TEL [FP5] project and its successors (TEL-ME-MOR [FP6], EDL [eContent*plus*], TEL*plus* [eContent*plus*, in negotiation], and EDLnet [eContent*plus*, in negotiation] are also inherently European and not national in that they are building an undertaking which involves every member state. The consortium created for the delivery of this project is deliberately focused on a core of libraries from just five Member States (Netherlands, Austria, Germany, the United Kingdom and France) but aims to deliver benefits to all contributors and users of the new European Digital Library and expand its scope in the second half of the project to include a wider range of national digital collections in a diverse and multilingual environment.

B 3.5 Centres of Competence

Over the past decade, there has been much investment in digitisation activities across Europe. One result of this has been the creation of many centres of competence on digitisation across the MLA sector.⁴ Since the Lund Action Plan of 2001, there has been considerable interest in the better coordination of digitisation activities across Europe, e.g. for helping to avoid duplication of effort. IMPACT provides an opportunity to focus attention on the practical challenges of developing technologies and workflows for the large-scale digitisation of texts and will play a leading role in informing EU-wide coordination in building the European Digital Library.

Wherever possible, IMPACT will work with other European centres of competence in digitisation to avoid the fragmentation and duplication of effort across Europe. Coordinating centres of competence forms a major part of the IMPACT vision. The project itself brings together partners from national and research libraries, research institutions and the commercial sector that themselves represent centres of competence with a vast knowledge and experience of large-scale text digitisation processes and technologies. One of the core attributes of the IMPACT consortium is that it provides a wide range of competencies, including significant expertise in research and development, in the production of software, in dissemination, and the practical application of technologies. As well as itself providing a good example of international and cross-sectoral co-operation on digitisation, the project will also co-operate with other centres of competence across Europe, e.g. through the provision of a well-supported and documented suite of tools and applications, through the development of benchmarking criteria for large-scale digitisation initiatives, and through the coordination of outreach and training activities. The project also plans to expand its scope by involving new partners in specific demonstration projects. In short, IMPACT will take a lead in helping to coordinate all activity in Europe related to the development of large-scale digitisation for textual materials.

⁴ For example, see the list at: <http://www.minervaeurope.org/competencecentre.htm>

B 3.6 Horizons

It is important to differentiate between short, medium and long-term impact of results in the following way:

- Results from (e.g.) surveys, studies and workshops that will have an immediate impact within the project itself and in participating organisations and, within 1 – 2 years, in other organisations once they become aware of the project through the dissemination activities of the Sub-Project *Capacity Building*.
- Technological development and eventual productisation within a three to five year time frame (i.e. development of an existing technical baseline resulting from work undertaken during the first half of this project)
- New insights and knowledge resulting from longer-term research strands within the project, where benefit may be realised beyond the life of the project itself.

In particular, in preparation of the annual rolling implementation plan, each proposed activity area will be categorised in terms of the prospective time-frame and impact analysis undertaken to help prioritise the themes for the next period.

B 3.7 Audiences

- National memory institutions can be reached through channels such as CENL (currently chaired by the project partner DNB) and The European Library Office (co-located with this project inside the Coordinating partner).
- Smaller collection-holders from the MLA community can be reached through national intermediaries.
- The research community can be engaged through platforms for which our research partners are themselves partially responsible (journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence and Pattern Recognition IJDAR International Journal on Document Analysis and Recognition, Computational Linguistics, Literary and linguistic Computing and conferences like ICDAR International Conference on Document Analysis and Recognition, DAS Workshop on Document Analysis Systems, DocEng the ACM Symposium on Document Engineering, COLING-ACL the annual conference of the Association for Computational Linguistics).
- Vendors and systems integrators can be targeted through their client-base within the MLA community.
- Innovative SMEs, both those already working in the collection-digitisation field and others for whom the EDL will provide new opportunities.