Digital libraries in FP7

Six European research projects

May 2008

The work programme

The first ICT Work Programme under FP7 defines the ICT research priorities for 2007-2008. Digital libraries research is part of Challenge 4, 'Digital Libraries and Content'.

Research objectives

The objectives for funding research on ICTs for digital libraries are twofold.

At medium term, projects should generate the conditions for the creation of **large-scale European-wide digital libraries** of cultural and scientific multi-format and multi-source digital objects, assisting communities of practice in the creative use of content in multilingual and multidisciplinary contexts, and based on:

- robust and scalable environments
- cost-effective digitisation processes
- semantic-based search facilities
- tools for preservation of digital content

At longer term, research should also explore **new approaches to digital preservation**, where advanced ICTs will have capacities such as:

- acting on high volumes of dynamic and volatile digital content (notably web content)
- safeguarding integrity, authenticity and accessibility over time
- keeping track of contexts (evolving meaning and usage)
- enabling automatic and self-organising preservation

Expected impact

- Unlock organisations' and people's ability to access digital content and to preserve it over time
- EU-wide massive digitisation and long term preservation of digital resources

The projects

The six projects presented in this brochure result from the first call for proposals under the ICT programme in FP7 (ICT Call 1, December 2006 - May 2007). There are four small or medium-scale focused research projects, one large-scale integrating project and one coordination action, which will be carried out by 64 participating organisations. The total amount of EU-funding is € 27 645 100. Start dates for the work were between November 2007 and March 2008.

IMPACT - Improving Access to Text

The IMPACT project aims to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitisation of the European cultural heritage.

more on page 6

LiWA - Living Web Archives

LiWA will develop and demonstrate web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long term interpretability.

more on page 7

PAPYRUS - Cultural and Historical Digital Libraries Dynamically Mined from News Archives

PAPYRUS aims at creating a cross-discipline digital library engine that allows for drawing content from one domain and making it available and understandable to the users of another.

more on page 8

PROTAGE - Preservation Organizations Using Tools in Agent Environments

The PROTAGE team will build and validate software agents for long-term digital preservation and access that can be integrated in existing and new preservation systems.

more on page 9

SHAMAN - Sustaining Heritage Access through Multivalent Archiving

This project will develop and test a next generation digital preservation framework including tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives.

more on page 10

Treble-CLEF - Evaluation, Best Practice and Collaboration for Multilingual Information Access

This Coordination Action supports the development and consolidation of expertise in the research area multilingual information access and disseminates this know-how to the application communities in the digital libraries field.

more on page 11

IMPACT

IMProving ACcess to Text

The IMPACT project aims to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitisation of the European cultural heritage.

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 on 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 experience of large-scale text digitisation processes and technologies. Within the project, they will 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 mass-digitisation programmes. IMPACT 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.

IMPACT's main goals are:

- Significantly **improve access** to historical text
- **Innovate OCR technology**
  - by exploring the challenges using different approaches, rather than from just one side
  - by developing cutting edge approaches such as collaborative correction
- Provide **innovative language technologies** to remove the historical language barrier
- Remove constraints to mass digitisation by providing **best practice guidance about the operational context for digitisation**

To ensure the inter-operability of the research results, IMPACT will define an overall technical architecture and monitor technical integration across all parts of the project. The project team will also deliver a coherent programme of dissemination, training and demonstration aimed at capacity-building in and beyond participating institutions. Particular attention will be made to addressing the needs of end-users and holders of collections of material in languages other than English.

Project facts:

Project type: Large-scale Integrating Project
Start date: 01/01/2008
Duration: 48 months
EU funding: € 11 500 000
Number of partners: 15
Project coordinator: National Library of the Netherlands, KB
Website: http://www.impact-project.eu/

LiWA
Living Web Archives

LiWA will develop and demonstrate web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long term interpretability.

The typical characteristics of Web content - variety of formats, high dynamics, volatility, interactivity and context-dependency - make adequate Web archiving a challenge. Research carried out under the LiWA project will look beyond the simple 'freezing' of pages, and develop tools to create 'Living Web Archives'.

'Living' here refers to:

- **long term interpretability** as the archive evolves and adapts over time,
- improved **archive fidelity** and authenticity by filtering out irrelevant information,
- captured content from a **wide variety of sources**.

To enhance archive fidelity and authenticity, LiWA plans to develop and test new methods based on content interpretation and intelligent pattern detection of traps and Web spam. The objective is to reduce the amount of fake content and help prioritise crawls by automatically detecting content of value.

To improve the integrity and temporal, structural and semantic coherence of Web archives, some work is dedicated to temporal Web archive construction. This serves the objective to significantly improve content positioning in time and (topic) space and will lay the foundations for fast and effective access to evolving Web content.

To facilitate archive interpretability, LiWA intends to apply methods for semantic and terminology extraction, able to detect and handle evolving semantics, interpretations of domain concepts and terminology. This shall contribute to the objective of preserving the usefulness, quality, and accessibility of Web archives over time.

For validating the LiWA approach, two demonstrator applications will be built on top of the LiWA services. These applications will focus on the social Web and on the special challenge of archiving audio-visual content.

The potential benefit of this research is twofold: Archiving institutions will be able to automatically archive higher volumes of dynamic and volatile digital content, resulting in a significant increase of preserved digital content. Archive users will benefit from the higher quality of archive content and improved search services.

Project facts:

Project type:	Small or medium-scale focused research project
Start date:	1 February 2008
Duration:	36 months
EU funding:	€ 2 682 400
Number of partners:	8
Project coordinator:	Leibniz University Hannover, L3S Research Center (Germany)
Website:	http://www.liwa-project.eu

PAPYRUS

Cultural and Historical Digital Libraries dynamically mined from News Archives

PAPYRUS aims at creating a cross-discipline digital library engine that allows for drawing content from one domain and making it available and understandable to the users of another.

Digital libraries generally provide electronic access for communities of users to information of their discipline. A new challenge to research is a digital library that would draw content from one domain and make it available to the users of another.

The PAPYRUS team approaches this challenge by introducing the concept of a cross-discipline digital library engine. This system will understand user queries in the context of a specific discipline, look for content in a domain alien to that discipline and return the results presented in a way useful and comprehensive from the perspective of the first discipline.

The use case chosen for this project is the recovery of historical content from digital news content.

The scientific and technological objectives of the project are:

- to advance the state of the art in **semantic multimedia analysis**, by introducing knowledge assisted methods which will take advantage of existing metadata and content structure models for the understanding of the source content;
- to propose context sensitive **query processing** methods, for the understanding of the user demands;
- to implement tools for automating the processof **knowledge mapping**, for corresponding concepts between the source content and the user queries;
- to develop **presentation techniques** for delivering search results in a form comprehensive to the targeted users.

To realise these objectives, PAPYRUS brings together expertise from research organisations with a focus on knowledge management, artificial intelligence and semantic multimedia analysis; experts in the history of science; two large European news producers; and a market-leader in search engines.

Project facts:

Project type:	Small or medium-scale focused research project
Start date:	1 March 2008
Duration:	36 months
EU funding:	€ 2 200 000
Number of partners:	9
Project coordinator:	Athens Technology Center (Greece)
Website:	http://www.ict-papyrus.eu/

PROTAGE

Preservation Organizations Using Tools in Agent Environments

This project will build and validate software agents for long-term digital preservation and access that can be integrated in existing and new preservation systems.

PROTAGE aims at addressing the problems created by the increasing volume and the heterogeneity of digital resources that have to be preserved, by developing tools allowing for more effective automation and self-reliance of preservation processes.

The PROTAGE solution is to link digital objects to long-term digital preservation processes by using software agent technology. Based on the latest research on digital preservation strategies and on autonomous systems, the project will build and validate **flexible and extensible software agents for long-term digital preservation and access** that can cooperate with and be integrated in existing and new preservation systems. Intended application areas for prototypes produced by the PROTAGE project cover submission of digital material as well as monitoring of preservation systems and transfer between repositories.

Tools developed by the PROTAGE project will:

- enable content producers to create and publish in a preservation-compatible manner,
- provide digital repositories with means of further automating the preservation processes,
- facilitate seamless interoperation between content providers, libraries and archives, and end-users throughout Europe.

Targeted end users are curators and digital content creators, including individuals managing their own digital collections. PROTAGE will use archive and library materials from the project partners for system and user tests and external stakeholders in further validation. The Swedish Centre of Competence for Long-term Preservation will ensure availability of results to a wider community of memory institutions. The industrial partners will use the results to develop commercial solutions.

Project facts:

Project type:	Small or medium-scale focused research project
Start date:	1 November 2007
Duration:	36 months
EU funding:	€ 2 021 900
Number of partners:	7
Project coordinator:	Riksarkivet (National Archives), Sweden
Website:	http://www.protage.eu/

SHAMAN

Sustaining Heritage Access through Multivalent Archiving

This project will develop and test a next generation digital preservation framework including tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives.

The aim of SHAMAN is to develop the framework for the next generation of long term (more than one century) digital preservation systems and tools. It includes the definition of a SHAMAN theory of preservation that integrates the analysis, ingestion, management, access to and reuse of information objects across distributed repositories. The data preservation capabilities offered will secure the authenticity and integrity of data objects through time.

The development work will be structured around four core components whose objectives can be described as follows:

- to establish an open distributed **resource management infrastructure framework** enabling grid-based resource integration, reflecting, refining and extending the OAIS model and taking advantage of the latest state of the art in virtualisation and distribution technologies from the fields of GRID computing, Federated Digital Libraries, and Persistent Archives;
- to develop and integrate technologies to support **contextual and multivalent archival and preservation processes** which are adapted and significantly extended from the fields of content and document Management and Information Systems;
- to develop and integrate technologies to support **semantic constraint-based collection management** to target one of the key challenges in automating one class of digital preservation core functions;
- to support the managing of future requirements by **securing interoperability with future environments** and maintaining essential properties of the preserved content.

Three prototypical applications will support trialling and validation in the following domains: i) scientific publishing in libraries and documents in governmental (parliamentary) archives, ii) digital objects used (eg CAD) in industrial design and engineering and iii) data resources used in e-Science applications.

SHAMAN's dissemination and exploitation plans aiming at actively foster outreach and take-up of results will be tailored according to the specific needs of the scientific / academic world and of industry users. SHAMAN's work will be naturally coordinated with other digital preservation European projects (CASPAR, PLANETS, DPE) as well as initiatives at national (DGrid, Germany) and international level (NDIIPP/NSF, US).

Project facts:

Project type:	Large-scale Integrating Project
Start date:	1 December 2007
Duration:	48 months
EU funding:	€ 8 398 300
Number of partners:	18
Project coordinator:	INMARK Estudios y Estrategias, S.A., Spain
Website:	http://www.shaman-ip.eu/

Treble-CLEF

Evaluation, Best Practice and Collaboration for Multilingual Information Access

This Coordination Action supports the development and consolidation of expertise in the research area multilingual information access and disseminates this know-how to the application communities in the digital libraries field.

The popularity of Internet and the consequent global availability of networked information sources and digital libraries have led to a strong demand for multilingual access and communication technologies. These technologies should support the timely and cost-effective provision of knowledge-intensive services for all members of linguistically and culturally diverse communities. This is particularly true in the multilingual setting of Europe. Despite recent research advances, there are still very few operational systems available, and these are limited to the most widely used languages. The Treble-CLEF project was launched to tackle the challenge how to best transfer the research results to a wider market place.

The project will build on and extend the results already achieved by the existing Cross Language Evaluation Forum. The aim is not only to support the development and consolidation of expertise in the multidisciplinary research area of multilingual information access but also to disseminate this know-how to the application communities. The specific target is the European digital library context.

Treble-CLEF will thus:

- support the annual **CLEF system evaluation campaigns** with tracks designed to meet the specific requirements of the user and application communities, and particular focus on user modeling, language-specific experimentation, and results presentation;
- launch a concerted action of **technology transfer and dissemination** of know-how, tools, resources and best practice guidelines through the organisation of workshops, tutorials and training sessions;
- encourage **community-building and collaboration** around this topic by providing a forum for the discussion of results and making the scientific data, experiments and results produced during the course of an evaluation campaign publicly available.

Project facts:

Project type:	Coordination Action
Start date:	1 January 2008
Duration:	24 months
EU funding:	€ 842 497
Number of partners:	7
Project coordinator:	Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Italy
Website:	http://www.trebleclef.eu/

DigiCult

The European Union is funding research that explores leading-edge information and communication technologies for accessing, experiencing and preserving cultural and scientific resources.

This programme (DigiCult, for short) is managed by the European Commission's unit 'Cultural heritage and technology enhanced learning'. The unit also supports research on how the use of ICTs can help make learning more efficient (technology-enhanced learning programme).

'Cultural heritage and technology enhanced learning' is part of the Directorate-General 'Information Society and Media', and one of the units of the Directorate 'Digital Content & Cognitive Systems' (Luxembourg).

For more information:

Website: http://cordis.europa.eu/fp7/ict/telearn-digicult/

Mailbox: infso-digicult@ec.europa.eu