EoI
Expression of Interest for an Integrated Project: European Web Archive

prepared by a Consortium consisting of
the National, State, or University Libraries of
Austria, Czech Republic, Denmark, Finland, France, Germany,
Italy, The Netherlands, Portugal, Sweden,

as well as companies and research centers at the
Vienna University of Technology (Austria),
Masaryk University in Brno (Czech Republic),
INCAD s.r.o., AiP Beroun s.r.o. (Czech Republic),
Center for Scientific Computing (Finland),
French National Institute for Research in Computer Science and Control INRIA (France),
Xyleme S.A. (France),
Fraunhofer Institute (Germany),
National Research Council CNR (Italy),
INESC-ID (Portugal),
University of Lisbon / LASIGE (Portugal),
Technical University of Kosice (Slovak Republic),
Intersoft a.s. (Slovak Republic),
University of Glasgow (UK),
Digital Preservation Coalition (UK and Ireland),

Abstract

This document represents an expression of interest for an *Integrated Project* following an invitation to submit expressions of interest to help prepare the first calls of FP6 (EOI.FP6.2002).

With the growing importance of the Web and its evolution from a technological playground to one of the core infrastructures, an "information mega-store" with tremendous diversity of information artefacts, awareness has risen for the pressing need to archive it as an entity, i.e. the documents, their structure and technology, as well as to use the information constituted by it. Numerous initiatives thus are being created, aiming at the collection of on-line publications, databases, or Web pages in general, be it delivery or deposit of documents, selectional or free harvesting, their preservation for the future, or analysis of the Web in terms of content, structure, and technology.

The ambitious goal of a European Web Archive not only provides an indispensable asset for our digital cultural heritage. It also represents a mirror of society and its needs, communities and their languages, of technology and market evolution, with far-reaching consequences for numerous application domains, such as administration (e-government, e-democracy).

In order to establish and benefit from such an archive, the integration both of the various projects addressing the same goals, as well as of the expertise from a variety of disciplines within the scope of an integrated project is of utmost importance to allow the information and complexity of the Web to be preserved, analyzed, and the knowledge it represents to be used.

Keywords: Digital Cultural Heritage, Preservation, Internet Archive, Digital Content, Information Retrieval, Knowledge Extraction, Data and Web Mining, Information Visualization, Internet Evolution

1 Motivation

Recent years have not only seen an incredible growth of the amount of information available on the Web, but also a shift of the Web from a platform for distributing information among IT-related persons to a general platform for all levels of society. It is being used as a source of information and entertainment, forms the basis for e-government, and e-commerce, has inspired new forms of art, and serves as a general platform for meeting and communicating via various discussion forums. It by now attracts and *involves a broad range of groups in our society*, from school children, professionals of various disciplines, up to seniors, all forming their communities on the Web, consuming and sharing experience, using it for transactions in different ways. At the same time, technology keeps advancing, providing new forms of interaction, new designs of portals, changing forms of interaction, thus shaping and educating society in various manners, influencing the use of language and the way conventional media are presented and used. The Web thus turns into a *mirror of society*, of the variety of *cultures* and *minorities* present, the advancement in *technology* accepted or rejected, the *content* considered relevant or insignificant, taking it beyond its mere existence as a collection of documents.

The recognition of the importance of such a core technology has led to an increasing awareness of the necessity to preserve it in archives, as well as to analyze the information it constitutes, providing a basis for analyzing and understanding the evolution of the medium, the society, and the technology in the past, as well as forming a basis for envisioning future trends and their prospects. This awareness led to the *creation of numerous initiatives* mostly on a national scale, developing specialized tools for the acquisition of Web Archives or their analysis. Information starting from 1996 is preserved in a limited way by some spearheading initiatives, while information before that date, the early days of the Web, as well as its forerunners, is already lost. The need to commence this task of a *European Web Archive* is pressing.

With the *Internet Archive*, the US has a private non-profit organization collecting and archiving, among other items, pages from the Web at a multi-terabyte scale. In Europe we find a number of initiatives, mostly on a national level, addressing these issues, having reached a significant level of expertise, using a variety of systems. Several National Libraries, in cooperation with partner institutions, have begun building different kinds of archives of their national Web spaces, collecting on-line journals, dissertations, or Web pages in general. Other groups are addressing important issues related to the maintenance of such archives e.g. the management of very large data stores and meta data requirements, as well as the preservation of digital material in general, covering issues such as media durability, file format conversion, emulation, and others. Yet another stronghold of expertise is present at the level of data analysis, having groups pursuing research in the range of technologies that can put this kind of archives at a level of use to the public that goes beyond surfing in past issues of the Web or browsing an archive of dissertations, providing content and structure analysis, vertical portals, and others.

However, we consider it important to *integrate these fragmented national initiatives*, in order to form a strong corpus being able to tackle the numerous challenges together in an efficient way, to find and develop state-of-the-art, consolidated solutions to the numerous challenges, and thus to build a distributed European Web Archive, *preserving our digital cultural heritage*, and benefiting from the wealth of knowledge constituted by such an archive. This is also reflected by the DELOS Report on *Digital Libraries : Future Directions for a European Research Programme*[1], which states that *"The grand challenge envisaged is the following: Establishment of an Initiative for an Integrated European Cultural Digital Library, which leads to the development of a comprehensive Digital Library of European history and cultural heritage."* By uniting existing initiatives, as well as by integrating a wide range of national archives, the basis for a coherent European Web Archive, distributed among the participating institutions, is formed. The need for an integrated European solution is pressing, as more and more institutions are currently realizing the importance as well as the potential offered by such archives, resulting in a heterogeneous, isolated landscape of archives with limited integrating value. Furthermore, the cooperation of a variety of IT disciplines, including among others, *databases, Web technologies, data and Web mining, digital preservation, knowledge management*, and *user interfaces*, to name but a few, is required to achieve such an enterprise.

While significant excellence in each of these disciplines exists at a variety of institutions within Europe, their bundling in the form of an *Integrated Project* is essential to the successful achievement of this vision of having and being able to use a European Web Archive, with the challenges in the dimensions of Web computing fostering and requiring significant scientific advances in each of the fields. Secondly, with the Web constituting a very heterogeneous and complex domain with respect to data acquisition, archiving, and analysis, the competence gained in handling these issues in the context of a European Web Archive will be applicable to numerous other, smaller-scale, and more controlled domains, such as the creation and preservation of company-internal Intranet archives, knowledge extraction and representation from heterogeneous data repositories, navigation within knowledge spaces, and others.

[1] `http://www.iei.pi.cnr.it/delos2/International/brainstorming.htm`

2 Approach

To achieve the ambitious goals underlying the European Web Archive, the cooperation of experts from a variety of disciplines are required, addressing the challenges at three core levels:

2.1 Acquisition of Information

The project aims at the preservation of born digital materials. Several strategies for data acquisition can be followed, each of them with important consequences for the archive, and complementing each other. Several approaches are currently explored by the various institutions, dealing with questions such as source selection vs. open collection, active collection vs. passive delivery, manual collection vs. free harvesting, snapshots vs. continuous harvesting vs. focused crawls, and others. The consequences of each decision need to be carefully evaluated, and best-practice guidelines as well as complementary tools are necessary to follow a comprehensive strategy.

- **Source Selection**: manual selection, open collection, semi-automatic selection of important resources, on-line journals, documents of government agencies
- **Data Collection**: client-based delivery, snapshots, focused crawls, acquisition of dynamic objects, interactive sites, session-filming, etc.

2.2 Archive Organization and Preservation

Apart from significant challenges with respect to the creation of such an archive, the amounts and types of data encountered call for new methods and technologies to be developed in order to maintain such archives, to provide access, that go beyond technologies known from and used for conventional cultural heritage initiatives. Techniques for processing and storing vast amounts of data, the need for replication and distribution to prevent accidental loss, as well as both short-term as well as long-term approaches to guarantee access to the stored information, pose significant technical challenges, requiring both the combination of existing as well as the development of new technology, with some of the core issues being:

- **Scalable Storage Technologies**: information-GRIDs, HDD-arrays, tape robots
- **Maintenance**: metadata generation, coding, storage, exchange and maintenance
- **Migration and Refresh**: storage media migration and refresh
- **Access Preservation**: system emulation, format conversion

2.3 Archive Navigation, Interoperability, and Information Mining

Apart from creating and maintaining such an archive, its value as an overwhelming source of information is to be made available in forms that go beyond the mere location of specific objects. Technology trends from the past can be extracted and used to predict the impact and evolution of future trends, the detection of communities on the Web provides an image of the evolution of society, the analysis of content and the evolution and use of language provide a deeper understanding of the needs of society and their changes, to name but a few. Yet, these kinds of analysis require the development and adaption of new means of organizing, exchanging, sharing, mining, and presenting information, making the implicit knowledge of the Web explicit, using a variety of technologies, ranging from Data Warehouses, via Natural Language Processing and Data Mining, to Soft Computing, with some of the core tasks being:

- **Information Retrieval**: indexing, search engine, audio and image retrieval, intelligent agent based off-line retrieval
- **Information Access**: new interfaces, interoperability, thematic portals, semantic web, ontologies
- **Information Mining**: topic detection and tracking, language usage and evolution
- **Technological Development**: technological evolution, geographic distribution, metropolitan media Internet, evolution of access support for people with special needs
- **Society Analysis**: communities, digital divide

3 Objectives and Required Results

The ambitious goals of such an integrated, distributed European Web Archive require an Integrated Project to combine the expertise from researchers and users of a variety of disciplines necessary for the development of the technologies required for the creation, maintenance, and usage of a European Web Archive. It requires a vertical integration covering the complete lifecycle of archived objects, ranging from information providers, via archical institutions, to the usage and exploitation of the knowledge represented by such an archive. It furthermore

calls for a broad basis as well as for a network of distributing the system, allowing other institutions to set up archival nodes, integrating both research, development, as well as take-up and training activities.

Yet the *applicability of the results* obtained within this project go beyond their immediate application area of the European Web Archive. The maintenance of large amounts of data, be it Web data, project data, electronic catalogs, is an eminent requirement in any medium-to-large enterprise, as is the archiving and preservation of relevant data. Technologies with respect to the analysis, retrieval, and visualization of knowledge, navigation within knowledge spaces and their interpretation are core requirements for the fostering of the information society. With the given scenario of a European Web Archive, these issues need to be addressed at their largest scale in a most volatile environment, requiring the development of technologies that can be applied in a range of related, smaller-scale fields.

To achieve these goals, such a project has to and can *build on the expertise and results already developed* by various national groups in the course of previous national, bilateral and international projects in their respective fields. The expected results of any such project can be briefly summarized as follows:

- **Archival System**: Development of a generic architecture for the delivery, deposit, harvesting, storing and retrieving of information from the Web. Development of components for that architecture, which shall form the basis of a European Web Archive Architecture, able to be used at a low cost by any participant institution. This architecture shall comprise also modules to provide access to these distributed archives in an integrated manner, support the maintenance of the collection as well as analytical services for the archive, and provide interoperability to not only the overall European Web Archive but also to other possible international networks.

- **Best Practice Reports and Standards**: Apart from a system for building a Web Archive, best practice reports and open inter-operation standards shall guide in the decisions to be made, concerning the type, operation mode, requirements etc. for the set-up, operation, and maintenance of any such archive. Experience from archive creation and usage models shall be used to help define a currently lacking legal framework for the archiving of on-line publications. Furthermore, lessons learned from the creation and maintenance of Web Archives can be used to provide guidelines for company-internal archives, governmental bodies, as well as archival institutions in general, with respect to digital content design, maintenance and archiving, as well as their long-term preservation.

- **Training Material**: Training material as well as supportive actions shall be provided in order to offer technical support and guidance, and to assist in the set-up and maintenance of additional archives to become part of the European Web Archive, and forming a cooperation infrastructure for an Internet e-archive community.

- **State-of-the-Art Research Results**: In the course of the Integrated Project, research results furthering the state-of-the-art in research in the respective fields, and that are applicable in related areas of information organization and distribution, preservation, visualizing and sharing knowledge will result, enforcing the strong position of the respective research groups.

- **European Web Archive**: The ultimate goal of the integrated project is the creation of a distributed, collaborative Archive of the European Web, accessible by everybody for exploration via commonly available Web interfaces, as well as a basis for research projects from a wide range of disciplines via special interfaces, supporting actions in the fields of e.g. socio-economic or cultural development.

4 European Activities

Numerous activities and research centers are addressing aspects of Web archives in their respective fields, currently funded by the various National Libraries, individual national initiatives, bilateral agreements, or on an European level. Contacts on a broader scale have already been established during a specialized workshop on Web archiving[2], and a core consortium comprising a large number of National Libraries, Research Institutions and Enterprises has been formed, resulting in a strong group of 27 core partners involved in the definition of the present Expression of Interest, representing outstanding expertise in the identified domains. Furthermore, the National Libraries members of the project will put all their efforts in disseminating it especially among the other members of the CENL - Conference of European National Libraries to allow for a wide-spread acceptance. Contacts with the European Commission on Preservation and Access (ECPA) have further been established, expressing their intention to expand their activities into the digital domain. Apart from that, there exist close ties to the respective non-european initiatives, such as the Internet Archive, guaranteeing and fostering international cooperation and allowing the European Web Archive to be positioned within a global cooperation of digital memory institutions.

The investments into such a European Web Archive project following the scope identified above, based on experience from previous initiatives in this field, can be roughly estimated in the range of 15-20 Mio Euro.

[2] *"What's next for Digital Deposit Libraries? Preserving On-Line Content for Future Generations"*, Workshop held at the 5th European Conference on Research and Advanced Technologies for Digital Libraries (ECDL), Darmstadt, Germany. 2001.

Below, we provide a list of members of the current proposing consortium of the EoI, organized by country, listing their respective area of expertise and interests.

Austria :

- Austrian National Library: *(Johanna Rachinger, rachinger@onb.ac.at; Alfred Schmidt, alfred.schmidt@onb.ac.at)*
  The Austrian National Library has been actively collecting on-line journals for some time, as well as initiated, in cooperation with the Vienna University of Technology, a pilot study for building the Austrian On-Line Archive.
- Vienna University of Technology: *(A Min Tjoa, tjoa@ifs.tuwien.ac.at; Andreas Rauber, rauber@ifs.tuwien.ac.at)*
  The Department of Software Technology and Interactive Systems is one of the key research institutes in the fields of Data Warehousing and Data Mining. It furthermore currently hosts the Austrian AOLA Web archive.

Czech Republic :

- National Library of the Czech Republic: *(Ludmila Celbova, ludmila.celbova@nkp.cz)*
  The NL has been actively engaged in Web resource issues (acquisition, preservation, access provision, legal aspects) for two years. It hosts the Czech Web Archive, developed in cooperation with the Masaryk University in Brno.
- Masaryk University: *(Miroslav Bartosek, bartosek@ics.muni.cz; Petr Zabicka, zabak@mzk.cz)*
  The Institute of Computer Science of Masaryk University has long been active in the fields of digital libraries, metadata (Dublin Core) and for the last two years it also develops metadata and Web harvesting tools for the Czech Web Archive.
- INCAD s.r.o.: *(Pavel Kocourek, kocourek@incad.cz)*
  Incad provides knowledge management applications, library management and manufacturing information systems, and in cooperation with the infrastructure provider CONERVA offers indexing, search and retrieval services.
- AiP Beroun s.r.o.: *(Karel Kucera, karel.kucera@aip.cz)*
  AiP Beroun is the producer of the Tornado fulltext search system optimized for extremely large databases, and used e.g. for the national bibliographies of the Czech Republic and Slovak Republic. It provides continuous fulltext indexing to structured and textual data in all European character sets.

Denmark :

- Royal Library: *(Birgit Henriksen, bnh@kb.dk)*
  The Royal Library has been archiving the static part of the Internet since 1998 as part of legal deposit of Internet material in Denmark and is active in the fields of Web harvesting and Web archiving since 1998. It is also participating in the Nordic Web Archive Project.
- Statsbiblioteket: *(Birte Christensen-Dalsgaard, bcd@statsbiblioteket.dk)*
  The Statsbiblioteket is investigating methods for building webarchives, new methods for building digital news and broadcast archives, as well as filming strategies in connection with webarchives, the creation of a European preservation strategy, as well as common software repositories.

Finland :

- National Library: *(Juha Hakala, juha.hakala@helsinki.fi)*
  The NL has been involved with harvesting of Web resources since 1997, first within the NEDLIB project, then as a partner in the Nordic Web Archive initiative. Using the NEDLIB harvester, the library has collected about 11 million files from the Finnish Web space. In NWA the library has developed some of the software modules needed for accessing the archived materials.
- Center for Scientific Computing: *(Mika Rissanen, mika.rissanen@csc.fi)*
  CSC has, in co-operation with the national library and other NEDLIB and NWA partners, built the NEDLIB harvester, which is used in numerous Web harvesting initiatives in Europe.

France :

- Bibliotheque Nationale de France *(Julien Masanes, julien.masanes@bnf.fr)*
  The French National Library is experimenting on focused harvesting and deep web archiving.
- INRIA *(Serge Abiteboul, Serge.Abiteboul@inria.fr)*
  INRIA is one of the largest computer science research institute in the world. It is one of the three W3C host institutions (with MIT and KEIO). The VERSO research group is working on efficient storage for huge quantities of XML data (hundreds of millions of pages), data acquisition strategies, query processing with indexing at the element level, change control with services such as query subscription and semantic data integration.
- XYLEME S.A.: *(Patrick Ferran, patrick.ferran@xyleme.com)*
  Xyleme provides technologies for intelligent crawling of Web information and for storing, indexing and managing very large volumes of XML data.

Germany :

- Fraunhofer Institute: *(Erich Neuhold, neuhold@ipsi.fhg.de; Ulrich Thiel, thiel@ipsi.fraunhofer.de)*
  The Institute for Integrated Publishing Systems (IPSI) at Fraunhofer Institute is engaged in several research projects in the fields of information and knowledge management, metadata and information structuring.

- Die Deutsche Bibliothek: *(Elisabeth Niggemann, niggemann@dbf.ddb.de)*
  Die Deutsche Bibliothek is collecting on-line dissertations since 1998. Furthermore DDB carried out a pilot project with publishers and other producers of on-line publications concerning the collection and long-term preservation of digital documents.
- Niedersächsische Staats- und Universitätsbibliothek Göttingen: *(Elmar Mittler, mittler@mail.sub.uni-goettingen.de)*
  The library is active in the fields of long-term archiving, subject garteways, and virtual libraries, and is further engaged in a number of DFG, EU, and international funded projects (e.g. EULER, Renardus, DIEPER, Math-Bib-Net Project, Meta-Lib Project, CARMEN, etc.)

Italy :

- Consiglio Nazionale delle Ricerche (CNR): *(Fabrizio Sebastiani, fabrizio@iei.pi.cnr.it)*
  CNR is actively involved in a number of large digital library projects, and has core expertise in the fields of digital library interoperability, personalization services, and text analysis.
- Biblioteca Nazionale, Centrale Firenze: *(Giovanni Bergamin, giovanni.bergamin@bncf.firenze.sbn.it)*
  The National Library is engaged in initiative for building a national Archive of the Web using the NEDLIB harvester.

The Netherlands :

- Koninklijke Bibliotheek: *(Trudi Noordermeer, trudi.noordermeer@kb.nl)*
  The KB has been leading the NEDLIB project. Main area of interest and activity include harvesting, selection policy and long term access.

Portugal :

- National Library of Portugal: *(Jose Borbinha, jose.borbinha@bn.pt)*
  The National Library is engaged in works relating to the acquisition of information (models for deposit and delivery of contents), interoperability (models, protocols and metadata), and preservation.
- INESC-ID: *(Alberto Silva, Alberto.Silva@inesc.pt)*,
  INESC-ID provides solutions in the fileds of metadata generation, coding, storage, exchange, maintenance, as well as interoperability.
- University of Lisbon - LASIGE: Large-Scale Information Systems Laboratory: *(Mario Gaspar Silva, mjs@di.fc.ul.pt)*
  LASIGE works in the fields of information acquisition, models for harvesting, indexing and retrieval.

Slovak Republic :

- Technical University of Kosice: *(Jan Paralic, paralic@tuke.sk)*
  The Technical University of Kosice is one of the leading institutions in the fields of knowledge-based systems, Web technologies, knowledge management, knowledge representation and knowledge modelling, and is actively engaged in several European Union Projects in these domains.
- Intersoft a.s.: *(Julius Kovac, office@intersoft.sk)*
  Intersoft provides IT development services and solutions for knowledge management, point-of-sale and mobile solutions.

Sweden :

- Royal National Library: *(Allan Arvidson, allan.arvidson@kb.se)*
  The royal National Libbrary has been collecting data from the Swedish national Web since 1996 and is maintaining the Kulturarw3 Web archive. It is also participating in the Nordic Web Archive Project.

United Kingdom and Ireland :

- University of Glasgow: *(Seamus Ross, S.Ross@hatii.arts.gla.ac.uk)*
  The Humanities Advanced Technology and Information Institute at the University of Glasgow is one of the core centers of expertise on preservation of data and the rescue of data from vantage hardware or software. It also heads the European Electronic Resource Preservation and Access Network of Excellence.
- Digital Preservation Coalition: *(Neil Beagrie, preservation@jisc.ac.uk)*
  The DPC is formed by a cross-sectoral alliance of members from the fields of libraries, archives, record offices, the Publishers Association, and Research Centers, and is active in the fields of distributed web archiving and harvesting, long-term preservation and storage technologies, selection policies and legal frameworks, best practice guidelines, development of value-added research and community web archive collections.

5 Conclusions

Our information society is increasingly relying on the existence of digital information for all aspects of its existence, ranging from e-government, via e-commerce to leisure activities, and is both shaping and being shaped by its technology. We are convinced, that these issues need to be addressed immediately in a comprehensive manner in the form of an Integrated Project in order to meet the challenges faced by our information society, ensuring its functioning, preserving its digital cultural heritage, and understanding its evolution in its many facets.