Introduction

The web, with its increasing functionalities is becoming one of the most important information and communication mediums. The number of digital resources available online are increasing exponentially. There is also a growing awareness that the web and its related sources are only tools and not finalised resources of high quality.

In all domains the Internet constitutes today one of the main information areas to disseminate, process and retrieve, and of course also to store information. The consequence is that there is now a huge corpus of information that needs to be preserved. In many cases the traditional resources do not exist any more or do not have the same value. The University of UCLA for example, has its university catalogue only available on the web. There exists no physical equivalent or even a digital file that is a "document" in the traditional sense (even in the form of a word-processing document). Another example is e-government, where the aims to provide 'users' (citizens) with direct and up-to-date electronic services require continued access through web-portals. Laws such as the UK Freedom of Information Act have a huge influence on the preservation and access to government information, demanding the continued availability of information that is on the web into the future. The Internet is not only a repository for information; it offers the added value of independence and wide dissemination for those making available the information.
The issues

Before consideration is given to the actual preservation of the material, thought must be
turned to the reasons why preservation is taking place. The future use and study of internet-based
information leaves a number of questions unanswered and, as Seamus Ross has recently stressed,
the problem is trying to understand how we will study cyberculture and what information must be
captured to ‘document our culture effectively’. ¹

Of course the dynamic nature of the Internet makes it very difficult to document and
preserve what often does not exist anymore and does not leave any trace after a few weeks or days.
The central role of cooperation in preserving this information cannot be underestimated.
Educational and training activities must jointly play a basic function in defining a framework of the
activities required for any level of preservation function.

There are a number of key points to be borne in mind. Firstly, the material on the web can
be split into two distinct categories: information that we can classify as a by-product of ‘real
communities’ (electronic publications as e-journals, e-books and web pages, electronic records in
any form, etc.); and information than can be classified as a by-product of ‘virtual communities’
(records of newsgroups, listservs, MUDS, chatrooms, etc.) which often pass directly into oblivion
like a conversation in the street, but are increasingly part of the routine transactions of institutions².
In addition, the special qualities of net based materials (audio, video, images, text, scripts) and
associated plug-ins and software, issues pertaining to copyright and data protection, and the
problems of selecting, accessioning and describing or cataloguing Internet derived materials all
place a great degree of tension upon preserving our cyberculture.

The possibility of building “a record of the Internet” (of course not to retain the enormous
amount of web-information available) requires the identification (more than in the non-web
electronic environment) of what, how and when we need to preserve. This should have specific
focus not only on the bits that constitute the objects, but also on the information able to identify the
behaviour of the rendering system, as Clifford Lynch, director of the Coalition for Networked
Information, has recently stressed with reference to the experiential digital objects³.

¹ S. Ross, “Cyberculture, cultural asset management, and ethnohistory. Preserving the process and
understanding the past’, Archivi & Computer, 2002, 1, p. 44.
² Ibid., p. 46.
³ Clifford Lynch, “Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central
Policies and guidelines are required, and indeed, some are in place.\textsuperscript{4} However, these are often focused more on the temporary harvesting of web material rather than its long-term preservation. In many e-government projects the requirements for best practice are implemented as national guidelines and establish some basic rules:

- to adopt and foster open solutions, which are platform independent and based on the use of XML syntax;
- to guarantee feasibility in terms of management by building scalable solutions (growth according to the real needs and to the financial resources available) and identifying more than one service manager;
- to foresee developments free of obligations to the software houses\textsuperscript{5}.

In many cases the projects “have not adopted a consistent methodology for selection, documentation, retention, access, and disposal of the data once it is collected”\textsuperscript{6}; too often attention is paid only to the content without context and without analysis of the processes related to the content creation and use.

We could conclude with Ross that the challenges connected to the web preservation cannot be solved with a centralised model, but should require “a distributed model, where organisations collaborate on tackling preservation...It leverages effort, ensures consistency of standards, and reduces redundancy of activity and duplication of content. The Internet is a multi-faceted environment in which collaboration, cooperation, and competition all play a key role, but it is only through collaboration that contemporary society we will be able to manage its preservation as a cultural asset and a social space”\textsuperscript{7}.

This seminar is aimed to offer a general perspective on the main existing projects, ensure a free discussion on the strategies and the solutions available and furnish a space for reinforcing collaborative links and distributed knowledge in new research areas.

**Dates and location**

The seminar, co-hosted by the University of Kerkira and ERPANET, will take place in Kerkira, from May 22\textsuperscript{nd} (2 p.m.) to 24\textsuperscript{th} (1 p.m.) at the University of Kerkira.

---


\textsuperscript{5} See for instance the technical requirements present in AIPA, ‘Servizio di cooperazione applicativa basato su event’, I quaderni, 1999, n. 3 in [http://www.aipa.it/servizi%5B3/pubblicazioni%5B5/quaderni%5B3/quaderni_3.pdf](http://www.aipa.it/servizi%5B3/pubblicazioni%5B5/quaderni%5B3/quaderni_3.pdf).

\textsuperscript{6} S. Ross, “Cyberculture…”, *ibid.*, p. 49. The article includes a list of the main projects in this domain.

\textsuperscript{7} *Ibid.*, p. 58.
Program outline

"Preservation of web material" is the focus of this seminar, with the aim of providing information on the latest research and offer the experiences of practical initiatives that are faced with the problem of preserving data, documents, online interaction, plug-ins – in short 'material' from the web. Specifically, the seminar intends to provide a detailed analysis of the main initiatives in the area of web archiving, and to discuss the long-term preservation of material obtained from Internet sources. Experts in this field will examine both technical and organisational aspects.

The seminar will focus on risk assessment, legal issues, selection/appraisal, deposit vs. harvesting, authenticity of web material, and preservation strategies. Metadata for the long-term preservation of web material, and guidelines and best practices to ensure long-term integrity and access to online digital material will also come out of the seminar.

The sessions, designed to be encourage discussion, will address the main topics of the seminar and will begin with an introduction and overview of the issues and highlight projects and definitions in this sector of digital preservation.

Session 1. Risk assessment and legal issues: the session will analyse copyright, privacy and data protection. These issues are complex to deal with in any digital environment, but in the case of web-archiving is even more difficult to define: who holds the copyrights in the web specifically in the case of dynamic systems, who is the author of a copyright protected work? How can these rights be verified and ensured in the long-term preservation? What are the implications of the copyright for the preservers? Does the act of reproduction for preservation purposes affect the copyright system? Privacy regulations also open a series of complex questions. The session will try to define the main risks and the basic legal questions preservers should be aware of in dealing with web material preservation.

Session 2. Appraisal and acquisition: as stressed in many documents related to the web-archiving, the questions on what and how to acquire and preserve, at what frequency, how to manage websites and make them available, are strictly related to the definition of different strategies and policies that can be followed. The session will try to give different examples and open a free discussion on the consequences involved in each strategy in terms of feasibility and cost of the long-term preservation.

Session 3. Best practices and policies for preserving and accessing the web material: is it possible to influence in the different domains (e-government, high education, scientific research, publishing, etc.) the creators of the web material? Which are the basic rules able to facilitate long-term preservation and strike the problem at the root? In this case, the seminar will provide some
examples in terms of methods and developing experience in the field of e-government, national libraries and archival sector.

Preservation of web material
Kerkira 22-24 May 2003

Draft programme

22 May 2003

15.00
General introduction: Andreas Rauber
Breakout session for general discussion

23 May 2003

09.30
Session 1: Appraisal and acquisition (what and how to preserve):
  – Direct access: Filip Boudrez (DAVID project)
  – Deposit: Annette Siegenthaler (Die Deutsche Bibliothek)

Session 2: Standards, rules and best practice for qualified web preservation
  – CINECA (Italy): an e-government experience
  – Minerva project: Pierluigi Feliciati (National Archives of Italy), a representative of the University of Patrasso

16.00
Breakout session for discussion on sessions 1 and 2

24 May 2003

09.00
Session 3: Risk assessment and legal issues (Why to preserve)
  – Risk assessment: Anne R. Kenney (PRISM project)
  – Andrew Charlesworth, (UKOLN, University of Bristol)

11.00
Breakout session for discussion on session 3

13.00
Conclusions (Erpadirectors)