

Cyberculture, Cultural Asset Management, and Ethnohistory –
Preserving the Process and Understanding the past.

Seamus Ross, Director HATII, University of Glasgow¹

Keynote Address at:

Preserving the present for the future - Strategies for the Internet

Blixensalen, The Royal Library, Copenhagen June 18-19 2001

(Programme available at: <http://www.kb.dk/tilmeld/konferencebeskrivelse.htm>)

Contents:

Introduction	1
Studying the Life on the Internet	5
What Content, Context, and Process?	6
More than Information: Imagination.....	9
Dynamic quality of web information.	11
Technology, IPR, and Lack of Policy?	13
Concluding Thoughts and Future Directions	15
Bibliography.....	18

Introduction

As a child I had spent three months each summer in a rustic seaside village without a television. I filled many of those TV-less evenings by reading articles from a battered, but much loved, copy of the twenty-nine volume 11th edition of the *Encyclopaedia Britannica* that my parents had bought for a couple of dollars at a yard sale. There seemed to be so much knowledge that had been gleaned from experimentation, the study of natural phenomena, the analysis of historical sources, and the investigation of remains of the past. Each article synthesised the work of many scholars and reflected many different methods of study. They all seemed to depend on or be linked in some way to many other articles. In the many decades that have passed since the 11th edition was produced our knowledge has exploded and indeed many of the ideas expressed in that edition—which could be described as a statement of human knowledge as it was then known—have been refined, revised, or even rejected. But the 11th edition had all the hallmarks that we expect from an information resource:

- ◆ its very character provided evidence of its authenticity and supported claims for the veracity of the information it contained (e.g. format of presentation, layout, presence of the initials of the author(s), it was backed by a major publisher and editorial system);
- ◆ it had internal consistency;

¹ see <http://www.hatii.arts.gla.ac.uk/>

- ◆ it appeared to be permanent (although by the time I read from it in the 1960s its pages had taken on a yellow-brown colour and become ‘a bit brittle’);
- ◆ it was complete;
- ◆ it was intelligible and accessible; and,
- ◆ the information that it contained was unique within it—that is there was no unnecessary repetition within the texts although there were many copies of the complete set distributed across many libraries and homes.

Although I, sadly, no longer have the luxury to spend summer days in the sea jumping the waves or on the shore building sandcastles and nights with a flashlight under the bedcovers reading from an encyclopaedia I remain fascinated by information and the interconnections between it. The attributes that made the *Encyclopaedia Britannica* a wonderful and seemly secure source of knowledge might provide us with pointers as to the attributes that we would expect of information that we might wish to select from the Internet for preservation.

The Internet is though more than just a massive digital library waiting to be harvested, processed, stored, and retrieved. It is an environment with high quality material on the virtual shelf alongside low quality and often illegal materials. We are not talking about retaining just these materials or the ‘text’, in a post-modern sense of the word. We now recognise the central importance of the social space, context, and interactivity that lie at the heart of the Internet. We consider the experience of using the Internet to be culturally defining. The physical and the virtual worlds are often contrasted, with the virtual world and its cyberculture viewed as uniquely different from ‘real-world culture’. While it is true there are characteristics of cyberculture that set it apart from more traditional measures of culture, the boundary between the two worlds has never been precise; the evolution of virtual social, information, and economic spaces has demonstrated this with remarkable clarity. Moreover cyberculture is not one culture it is many—it is the world of the scholar, the average citizen, and the deviant. The Web is a cultural asset that we must manage and it is its intellectual capital that we must selectively harvest and pass to future generations.

Before we begin to tackle the problem of digital preservation of the content of the Internet and in particular that of the Web we must tackle the problem of objectives. Why do we wish to preserve Internet-based information and to what purposes will it be put in the future? The problem is how will we study cyberculture and its by-products, and what information will we need to capture from the Internet to document our culture effectively. This is a problem of selection and one that most heritage institutions, such as libraries and archives, have struggled to address in other contexts. For example, few public and research libraries in the United Kingdom preserve copies of *The*

Sun newspaper, although it has a daily circulation of around 3.5 million². Many, however, keep copies of *The Times* although its daily readership is less than a fifth that of *The Sun*. Why is one paper retained rather than the other? One answer is that Librarians' agree that *The Times* is for the United Kingdom a paper of 'record'. But 'a record' of what: political opinion, news reporting, social commentary, sports, or all these domains. It is though neither a record of popular attitudes, beliefs, and activity nor of the way the national and international events are seen by the masses. From the point of view of popular culture *The Sun* provides a very rich window on the British and one that future historians would find invaluable. Popular culture has always been under represented in the documentation of the past as archaeologists and historians increasingly recognised during the second half of the 20th Century. In considering retaining the Internet we must ensure that we balance our collecting strategies so that we capture all the facets of our culture whether that is popular or high culture.

The dynamic nature of the Internet makes it difficult to document and often what you capture is not there in the future, even a few days later, or as in the case of Internet news services it may be gone within a matter of minutes or hours.³ This fluidity makes the scientific study of the Internet difficult. The experiments necessary to study it can rarely be repeated and the information it contains and its shape is always changing. Herein lies a parallel between the study of the Internet and its cyberculture and the work of anthropologists documenting 'real-world cultures.'

In considering the issue of preserving the cultural assets that exist in cyberspace we need to examine:

- ◆ the unique qualities of cyberculture and net-based e-materials (e.g. web pages, images, text);
- ◆ the problems of selecting (e.g. whether based on prioritised criteria or harvest everything), accessioning, and describing or cataloguing Internet derived e-materials (e.g. web pages);
- ◆ legal issues (e.g. copyright, privacy, data protection, national security, trademarks, patents);
- ◆ the growing complexity of collecting the web (e.g. caused by the increasing use of dynamic web pages served from underlying databases);
- ◆ the difficulties inherent in preserving and providing long-term access to and ensuring the usability of Internet-derived e-materials (e.g. the fact that some web materials are only accessible or usable in conjunction with particular software applications, popularly referred to as plug-ins); and,

² For *The Sun* average daily circulation between February and July 2001 was 3,487,015, whereas during the same period *The Times* had an average daily circulation of 710,709. See <http://www.abc.org.uk>

³ It is probably worthy of note, though, that in the case of news sites this is really no different than news papers that publish different daily editions (whether these vary depending upon time of the day or region).

- ◆ the need for methods for studying these materials where they can be preserved (e.g. who will be allowed access to Internet archives, what kind of uses of these archives will be appropriate, how will it be monitored and controlled).

What is different about cyberculture and Internet-based information resources from traditional materials? Essentially two kinds of information exist in the net-based environment:

- (a) information that we can classify as a by-product of 'real communities'; and,
- (b) information that can be classified as a by-product of virtual communities.⁴

The by-products of real-world communities include electronic publications such as e-journals, e-books and web pages, and the electronic records of institutions and their transactions, such as databases, the emails exchanged by Heads of State and their ministers and other Heads of State, which private networks or the Internet have been used to distribute. The by-products of virtual communities also include emails and web pages, but they include the records of newsgroups, listservs, MUDs, and chatrooms. Some of these latter 'by-products', like conversations between two strangers in the street, go unrecorded (e.g. chatroom interactions). As time goes on this distinction between the real and the virtual will become far more complex and much more blurred; see for example Julian Dibbell's troubling discussion of a *Rape in Cyberspace* (1996 and 1998) or Donath's examination of identity and deception in virtual communities (1999) as evidence of this.

It would be possible to focus on the by-products of real communities as the main emphasis of this discussion, but examining the interests and attitudes of contemporary historians suggests that future historians are going to be more interested in the social phenomenon of the Internet and its communication environment than in the electronic publications of scholars that were made available on it. Several years ago I mused that 'the long-term influence of networking on scholarship will depend in part on the increase in the quantity, diversity and quality of information resources available in digital form.' There is now a tremendous amount of information available on the web, much of which is of little use to contemporary academics. Certainly, the digital medium provides publication opportunities that print could never support, such as facilities to use a variety of data types (e.g. images or audio) and structures (e.g. databases) within a single scholarly work, to distribute that material widely, and to reanalyse dynamically data sets used in support of scholarly arguments. Some scholars have responded by taking advantage of these rich opportunities. At the same time the general public, and public and private institutions have recognised the Internet and in particular the web as a mechanism to distribute and provide access to information and as an environment in which to interact with other users whether these are local or

⁴ In both instances the material may have been born digital, but in the first it need not have been.

remote. The questions that we will find most exciting will be those that relate to how people used the Internet and its web-resources, what impact it had on shaping events, and what views it can offer us of popular culture.

Studying the Life on the Internet

In what is now the classic study of the Internet, *Life on the Screen*, Sherry Turkle described its social environment.. Her tale is though one of geeks, social misfits, and deviants whose lives became absorbed by the 'text-based' environment and who used the virtual world to create a life for themselves in the (apparent) absence of a real world life. One difficulty with her study is that she describes her data collection methods and experiments poorly and her conclusions prove difficult to verify. For the past two years in a seminar for honours and postgraduate students at HATII (at the University of Glasgow) my students and I have attempted to repeat her experiments and conducted new ones into identity, gender, ethnicity, and communication. In each of these years our conclusions have been different, although our experiments have been the same.⁵ This year when we compared the results of our two years worth of experiments with Turkle's study and considered the various explanations for the variance between the results the most plausible explanation proved the simplest one: the variance reflects the dynamic nature of the Internet, the information resources that it holds, and the natural variation in uses and user practices. What emerges is that Turkle's study, like so many other Internet studies, has become an historical text, which captured a fluid phenomenon at moment in its history; it is akin to an ethnographic study. At this point in the history of the Internet it is probably worth recognising the value of the ethnographic work as a method for recording and understanding the social practices and environment it engenders. Ethnography may provide us with the only way of charting cyberculture and net-based communities.

If we wished to build a record of the Internet that would allow us to repeat experiments such as those that Turkle or my students conducted what would we need to preserve:

- (1) the content;
- (2) the interconnections of knowledge as these linkages tell us much about how the general public, contemporary scholars, and institutions thought their knowledge was inter-related;
- (3) the interactivity of the Internet/Web itself;
- (4) the immediacy of the net-based environment;
- (5) the modes of access and a record of how they are changing (e.g. the transition from a 'text-based' to an 'image-based' environment);

⁵ A future article will pull together this work and explain the experimental methods that we used and the results of the experiments.

- (6) snapshots of how it is being used either as comprehensive session logs which enable one to replay a full user interaction or as video sequences;
- (7) examples of the hardware or sufficient data to make it feasible to emulate it; and,
- (8) the software applications themselves.

To make this happen we need to create:

- (1) mechanisms that enable self-documenting functionality and context—that is a facility to generate the metadata necessary to retain the information in its current and functional form;
- (2) large, and probably collaborative, data repositories;
- (3) virtual environments; and,
- (4) simulations of the way that all this information is used.

What Content, Context, and Process?

Brewster Kahle reported that in 1996 there were '50 million web pages with the average page online for only 75 days' (1997). In November 2000 I cited a figure that reported that there were one billion pages of information on the web (Ross 2000), but by March of 2001 there were some three billion pages available. Of course, arriving at precise estimates is complicated by the fact that the web as seen by many users and search engines is far from a complete view of the terrain. Many documents or web resources lie behind 'firewalls' and password protected sites, some resources exist as records in databases, and still others in file formats that the current generation of search engines cannot harvest.

There really is no way that we can expect to retain all this information for posterity. In the past we never expected to retain everything, and one might argue, although other archivists have suggested that this would be a error, that there are aspects of our 'web- or cyber-' culture (e.g. pornographic and 'hate' websites) that we might not wish to preserve for posterity. Although we tend to see preservation of the contents of the web as a major problem, it is only the most obvious manifestation of the information explosion of the last twenty-five years. For example, in 1995 Jones reported that the typical large corporation had 258 million pages of information—product manuals, company accounts, and marketing materials (Jones 1995). None of these firms could retain all this material for posterity and few would wish to do so. Many do have record retention strategies to ensure that 'necessary' records are retained.⁶ The problem has always been deciding what information is worth keeping, what will be useful to provide the future with a view of our society,

⁶ 'Necessary' could include legal requirements to retain certain materials, retention of records to support the organisation in case of litigation, or the retention of records that can provide a foundation for organisational memory.

and what strategies we can put in place so that we avoid passing our 'cultural myopia' to the future.

Private and national institutions in an increasing number of countries are taking snapshots of the web (see Table 1). So far these projects have not adopted a consistent methodology for selection, documentation, retention, access, and disposal of the data once it is collected. Many of these projects use different models for selecting material for 'web archiving'. Some such as Kulturarw³ (Sweden) attempt to harvest all sites within a particular domain⁷ whereas others such as Pandora (Australia) use a suite of selection criteria defined by subject specialists (Relf 1999). Among the current or soon to start web-archiving Projects are the following:⁸

Project Title	URL (valid as of 7/6/2001)
Alexa.com	http://www.alexacom/help/webmasters/request_bot.html
EVA	http://www.lib.helsinki.fi/eva/english.html
Kulturarw ³ Heritage Project	http://kulturarw3.kb.se/html/kulturarw3.eng.html
NEDLIB	http://www.kb.nl/nedlib
Nordic Web Index and Nordic Web Archive	http://www.lib.helsinki.fi/finelib/kopenhamn/hakala2/index.htm
PANDORA	http://pandora.nla.gov.au/
The Internet Archive	http://www.archive.org/about/index.html

These initiatives are at varying stages of development, some are still pilots and others are well advanced. The Minerva Project conducted by the Library of Congress, for instance, has just passed the prototype stage (Arms, et.al. 2001). As of March 2001 the 43 Terabytes held by the well advanced *Internet Archive* included some four billion pages, sixteen million Usenet postings (1996-8 and 2000-1), and 360 movies.⁹ In 2001, the French Government expressed its concern about the way the web had changed the mechanisms by which French culture was being created, presented, and not necessary preserved by passing a law that would require every French web page to be archived. While the primary obligation would fall to the publisher of the

⁷ This statement simplifies the collecting practices because the Kulturarw3 collects all websites ending in .se and those that end in .com, .org and .net which reside on servers based in Sweden, and a number of other addresses such as .nu. See <http://kulturarw3.kb.se/html/projectdescription.html>

⁸ It may also be worthy of mention that there are a growing number of companies that are developing products and services to assist other companies in preserving the web, such as Cambridge Computer Services Inc., (http://www.camcom.com/html/document_archiving_and_electro.html). Also of interest is OCLC's Electronic Archiving Project. Although not strictly a web archiving initiative, it has the potential and the infrastructure backing to be extended in this direction (<http://www.oclc.org/oclc/press/970127a.htm>).

⁹ <http://www.archive.org> (19/06/01)

website, the French law grants the national library and the national audio-visual institute the right to harvest the 'entire French web'.¹⁰ Implementation of the law will take some planning as issues such as what constitutes the 'French Web', technological problems with particular kinds of materials, and legal issues (e.g. whether software can be harvested), remain to be addressed. As the Minerva Project found, harvesting websites is fraught with difficulties from problems with formats to availability to databases (Arms, et.al. 2001).

Preserving objects or information without context is though meaningless. To take a poignant parallel, had Erwin Black not seen the IBM Hollerith D-11 Card Sorting Machine in the context of the Holocaust Museum in Washington DC he might never have wondered about the role that it and the corporation that made it played in enabling the Nazi's to bring about the Holocaust (2001, 11). Seeing the machine in the context of the Museum led Black on a quest. It resulted in his establishing a relationship between the machine and the collection, analysis, and management of information that lay behind the horrific events of the Holocaust.

As much as context is key to interpreting the meaning of information, so is process. The great focus on the preservation of electronic records, those records that document our society and the actions of its individuals and institutions, has been in such areas as ensuring that suitable metadata are created and linked to the records to ensure that they can be reconstituted, contextualised, and authenticated. The focus has been on ensuring that sufficient metadata exists so the whole process by which particular records were used can be recreated. One of the five core drawbacks¹¹ of Black's study stems from weaknesses in the description of the Hollerith technologies—the study does not bring to life the sensitivity and power of the technology. As a result the description leaves the process opaque. The whole way the information was processed and the technology made to work was at the heart of the horrific story he was telling. What is evident is that we can piece together the process by which information was used provided we have the content, suitable metadata, and the original machines, but without these it is nearly impossible to reconstitute the process. As a result we have an incomplete view (Ross 1998). The need to create a context and process environment for the information we harvest will be a major obstacle to the meaningful retention of the Internet, the web-based documents (e.g. texts, images) it holds, and the interactivity it enables.

¹⁰ <http://www.europemedia.net/shownews.asp?ArticleID=4075>

¹¹ The others are not discussed here.

More than Information: Imagination

The other problem is that the Internet is more than the sum of the information available from it. Five years ago in *Modernity at Large*, the University of Chicago anthropologist Arjun Appadurai proposed a new method for studying cultural globalization (1996). Building on Benedict Anderson's concept of the 'imagined community' (1983), Appadurai's groundbreaking study argued that imagination has become a form of social practice that underpins modern societies and Internet cultures. It gives them a sense of place, community, and time.

Imagination provides mechanisms to enable us to create, and even fabricate, identities, build communities, and in the view of Appadurai and Anderson generate alternatives to the nation-state. It has, therefore, become a pervasive and powerful social force. As Appadurai argues it plays an increasingly significant role in the lives of individuals as mass media, from popular magazines to broadcasting, offer us a rich array of possible lives against which to view, measure, and contextualise our own. What has happened is that fantasizing has become social practice for people in many societies; it provides us with the opportunity to give context, place, and possibility to our lives. In cyberspace we experience this and the other prospects imagination offers in a variety of ways. Such cyberspaces as MUDs and chatrooms are all fantasy worlds; in the HATII seminar 'Investigating Cyberspace' we repeatedly observed the critical role that imagination played in communication and in allowing us to fabricate and situate our place in the virtual social spaces in which we participated. The Internet enables individuals to share experiences and create social bonds through chatrooms, newsgroups, multi-users environments, and webrings. They enable us to build 'imaginary communities' that take on social and cultural fabric, but remain rooted in fantasy until they eventually become 'real'.

As most of these communities remain in 2000-1 'text-based' (in the traditional sense of the word text) they often cut across gender, cultural differences, ethnicity, and race to create new kinds of social groups and spaces. In addition to depending upon imagination they also rely on our ability to trust or at least suspend disbelief (see below). Just as in the real world individuals and groups seek to reproduce, strengthen, and promulgate their cultural identities they also seek to do these things in the virtual world. So while these communities are fragile and transient, and their interactions are often inadequately documented or unrecorded, there is an increasingly iterative relationship between Internet communities and the real world (e.g. consider the interplay between cyber-based disembodiment and body piercing, and that between virtual gender swapping and cross-dressing and trans-sexuality).

This anthropology of space and community is very time bound and its study is not dependent upon merely preserving the web pages with their ASCII text, their images (gifs, jpegs, or pngs), their audio, and their virtual reality representations. It requires that the inter-relationship between space, context, and imagination be retained. This is all about the 'text' in the broadest sense of the word. In an investigation of a co-related problem of audience research in the television environment, Virginia Nightingale argued the text cannot be separated from the vision that the audience creates of it (1996). As Nightingale reminds us this focus on context by electronic specialists is not new in the study of information or society. The formative anthropologist, Malinowski concluded from studying the Trobriand Islanders that 'not only the text, but the context, the 'whole nature of the performance' (its quality and timing) and its 'private ownership' contributed to its meaning' (Malinowski 1954). Studies of the Internet and in particular the web, chat-rooms, and MUDs bring us to the same conclusion: we cannot separate these environments and associated information from their use and a record of their users if we expect to reconstruct and understand them in the future.

Virtual communities, not unlike their real counterparts, represent shared systems of values, meanings, standards, and behaviours. Where individuals evolve a sense of community whether in the real or the virtual world they evolve mechanisms to protect their community and its collective identity. The sense of place and of belonging is consolidated by the evolution of guidelines and protocols of behaviour and methods to enforce them. In other words, communities coalesce around practices, rules, and symbols which members (or 'insiders') learn, follow (or even move to reject), and can certainly recognise.¹² These practices, rules, and symbols also serve the function of making it easy to identify outsiders. Not unlike real communities, virtual ones often become exclusionary and even isolationist. As Curtis reported in a study of 'MUD-life', 'The participants slowly came to consensus about a common (private) language, about appropriate standards of behaviour, and about the social roles of various public areas' (1996). Other studies of virtual communities have come to similar conclusions (Reid 1996 & 1999; see essays in Holeyton 1998 and in Smith and Kollock 1999). One of the reasons that communication is controlled by certain rules is that most of these communities are based on a very fragile balance of trust, which itself is closely related to imagination. The by-products of these virtual communities form the text that will enable us to understand virtual communities and the experiences of cyberspace. The challenge will be how to document them. Should we:

¹² It is worth remembering that it is possible to belong to multiple virtual communities using one or multiple identities, and a range of different cultures.

- (a) attempt to capture the whole context so that we can actually or virtually reconstruct these communities;
- (b) take snapshots from which future users will be able to 'establish a view' of cyberculture(s); or,
- (c) encourage the development of a new discipline or practice of web-ethnography supported by anthropological methods for conducting this work.

Dynamic quality of web information.

If we focus on the 'information' explosion itself for a moment we will all recognise that the web has changed the landscape of information creation, provision, and use. It is, though, a culmination of a chain of developments that created an environment ready for the possibilities it offers.

The contextual controls on the production of books and magazines meant that the process of selection of material for addition to a library's collection and eventual preservation was part of a chain of activity. This contained or restricted the amount of information available in print at anyone time. The reduction in the costs of production, changes in the processes of production, and the increasing ubiquitousness of methods of production, first with the advent of the photocopier and then with the laser printer, began to change this model. Publishers encountered economic challenges (or opportunities) and responded by shifting their publication models. For example, some moved from selling many copies of a few titles to a mixed strategy in which they did that and they began selling a small number of copies of many titles. This publication environment made many realise that they too could be authors of published and their works distributed in multiple copies. These two shifts flooded the market place and created an initial problem for libraries: how to respond to the explosion in available information resources. The advent of the web and the widespread realisation that we can all be authors and publishers, has exacerbated this problem still further.

The scale of the e-revolution is awesome. In January 2001 NASA reported that it had some 1.9 million web pages online. Staff at NASA estimated that if it took 10 seconds to call up each one and take a snapshot of it, the effort to create a 'moment in time view' of NASA's entire web presence would take staff 231 days working around the clock (*Federal Computer Week*, January 2001). Technology could be brought to the rescue here because, as we have seen, it is feasible to build intelligent agents that can harvest information, but we are a long way off intelligent agents that can document the information they harvest as effectively as humans. Of course in its time and effort projections NASA did not include the time that would be necessary to create contextual and administrative metadata because their estimate was a response

to a request from the National Archives for a snapshot of government websites (see below).

Historians have long noted the value of marginal notes on documents, such as routing lists which members of staff initial when documents pass across their desks (often taken as evidence that a particular person saw a particular document). In most computer-based environments these have proved difficult to preserve with the exception, of course, of transaction databases in which audit trails have been incorporated. Most web-preservation strategies focus on content without the context and evidence of use. As is evident in the case of the Hollerith machines and the Holocaust the 'cold mechanical sorting, selection, refinement, and extraction' of information that these machines made possible was the key to the process, but it was not sufficient. In the Internet world to understand process we will need to preserve examples of machines, user access logs, and evidence as to how users used these resources. Profiling the user in richer and richer ways and linking these profiles to the web resources themselves will be increasingly important if we are to make it possible for future historians to reconstruct the process by which the Internet and the information that it carried was used.

We must bear in mind that we cannot keep everything. There are a number of reasons for this:

- (1) it would be impossible to document it sufficiently to make it worthwhile doing so;
- (2) the information collected would be hugely redundant (Redundant information uses up scarce resources of cataloguers, computer storage space, and tends to enhance the perception of the importance of some information at the expense of other material.); and,
- (3) without the benefit of a contemporary view of what is significant it will be a very complex process for future researchers to separate the significant from the trivial—to differentiate the usefulness of the works by the Seneca's from that by the Sidney Sheldon's. Many current web preservation strategies will accumulate from the Internet 'the trivial with the profound, the erroneous with the factual, the fundamental with the general, and the simple with the complex' (Ross 1995, xvi).

There are those who argue that we can afford to be very selective in what we preserve because they wonder 'who will be able to use all this information anyway'. We must not assume that researchers in the future will be constrained in their ability to use digital archives by our current generation of technology. The web is a massive digital resource and we currently have crude tools to access and manipulate the information it holds. These are getting better all the time. Future researchers, whether scholars or the general

public, will have an increasingly rich array of tools to help them to discover meaning and information in the digital content that we preserve.

If we select material from the web for preservation we must ensure that the data have the same internal consistency that we expect from printed sources, such as my beloved 11th edition of the *Encyclopaedia Britannica*. The information selected must be:

- ◆ authentic and capable of authentication;
- ◆ accurate;
- ◆ internally consistent;
- ◆ complete;
- ◆ intelligible and accessible; and
- ◆ and permanent.

In addition it must be possible to make it functional—that is to render the information whether web pages or databases in a usable way—and the material must be unique. There is little point in preserving multiple copies of the same digital information since one of the great strengths of digital information is the ease with which it can be copied and transmitted across space and eventually time. Of course, we need to keep not merely the information, but also the technical infrastructure (e.g. the application plug-ins, such as image viewers) to access the information.

Technology, IPR, and Lack of Policy?

Preservation of our digital heritage, whether the web, e-records or e-documents, requires we address many technical obstacles. *Changing Trains at Wigan* describes many of these technical challenges and suggests ways that they might be addressed (Ross 2000). It notes that the most promising approaches will depend on new research in the areas of Magnetic Force Microscopy, binary retargetable code, and emulation. There are numerous other authors who have also looked at the technical issues and as a result we shall not labour these here.

Overcoming the problems of hardware and software obsolescence, media degradation, support, and documentation makes the preservation of electronic resources expensive. These costs do not decrease with time. Electronic resources require continual attention and, left unattended, quickly become inaccessible. To assume that material in electronic form will be available for future researchers is to underestimate the economic cost of migrating records from one type of media to another, the problems of suitably documenting them, and the difficulties of creating virtual environments where the content can be reused (e.g. do you have the correct version of the plug-ins?).

Moreover such legal instruments as copyright and data protection may make it impossible to preserve a lot of the information that is available on the web.

To download and retain, even freely available, information from the web may infringe the IPR rights of an individual or organisation. Tim Slagle in a message posted to *RISKS-LIST: Risks Forum Digest*,¹³ wondered, in 1997, whether the *Internet Archive*

‘could have the largest archive of violated copyrights in the world’[...although] ‘I don’t really have the interest or resources to pursue an infringement suit, but I bet someone else will (like an artist, publisher, or software company) will, especially if the Internet Archive starts allowing access to their data.’

Perhaps there is much truth in Slagle’s line of argument, but the legal issues associated with IPR may not be the only ones that will create difficulties for Internet archiving projects. In some countries data protection or privacy regulations may have an impact on web archiving practices, or at least a restrictive impact on who can access the digital archives and what they can do with them. The risks of infringing patents and trademarks are other possible obstacles.

What we must avoid is unplanned or poorly designed responses. The United States recently produced a wonderful example of such a situation. On January 12th 2001, eight days before President Clinton left office, Lewis Bellardo Deputy Director of the National Archives sent a *Memorandum To Chief Information Officers* of Federal Agencies asking them to take a snapshot of their agencies public websites on or before 20 January 2001 (see the NASA case above).¹⁴ Intranet and restricted materials were specifically excluded.¹⁵ An accompanying Guidelines document stated that ‘The snapshot needs to include all of the documents available to the public that are located on the agency’s web server(s)’ (*NARA Guidelines 2001*). The storage format that NARA required agencies adopt for this exercise depended upon media types and data representation standards that were widely recognised to be out-of-date .

What the document displayed was a remarkably limited understanding of the technology that was currently in use within US Government agencies. *Federal Computer Week* reported that: NARA’s call for snapshots of all public agency Websites is widely seen by Federal Webmasters as ‘an undoable request’. One Webmaster reported, ‘We want to comply with all of the rules, but lately it seems like more of the rules are being made by people who really don’t understand the Web very well’ (Matthews 2001).

Few organisations have successfully tackled the issue of even preserving their own websites on a local level. The European Commission, for instance, is (in 2001) actively working to develop a Website Electronic Records Management

¹³ 10 January 1997, Volume 18: Issue 75, <http://www.tao.ca/wind/rre/0019.html>

¹⁴ I am grateful to Charles Dollar for calling my attention to this example.

¹⁵ To be fair the memo did note that developing strategies for archiving websites was a priority and that the National Archives was working with federal agencies to make this happen.

(WERM) policy. There are many strategies that have been or are being developed at national level elsewhere, such as the policy produced by the *National Archives of Australia, Archiving Web Resources: Policy*.¹⁶ There is room for more, but more importantly there is a need for buy-in to these strategies on an international basis. Most companies are acting to retain their websites as part of their corporate records management strategies and their activities need to be factored into any moves to preserve the web if duplication of effort is to be avoided.

The preservation of the web is a layered activity with individual organisations putting in place policies and mechanisms to ensure that their web-based electronic records are preserved, and umbrella activities harvesting the web as a whole to ensure that as much as possible of the content that is currently accessible is preserved for the future. Strategic activity should leverage this layering rather than creating redundancy and inconsistencies in the material that is collected and preserved.

Few of the web archiving projects seem bothered about the accuracy, intelligibility, significance, or representativeness of the information that they are harvesting. Many focus on harvesting rather than selecting (Pandora is an exception in this). This seems to me likely to pose a major difficulty. One of the hallmarks of libraries is that they have struggle (although not always succeeded) to collect, make accessible, and preserve verifiable content.

Concluding Thoughts and Future Directions

A guiding principle for web preservation must be *fitness for purpose* and in this regard we must ask what will researchers of the future want and what can we expect to be able to give them. If a primary aim of preservation is to ensure that records of contemporary decision-making processes are preserved then the approach to preservation will be very different than if we expect to retain *the social space that is the Internet*. On the other hand we must not lose sight of the fact that web pages and web-accessible resources are by-products of the world of cyberspace and the individual texts while enticing will not be the only subject of future study. By way of comparison we might consider that, while philosophers examine the ideas expressed in Plato's *Symposium*, historians study the same document as an ethnographic text to understand the workings of the Greek Symposium.

If we are only interested in the information content of web pages and their context then the preservation challenges are not much different from those we face with any other kind of electronic resource (including electronic records). This strategy produces a patchy, decontextualised and incomplete record.

¹⁶ http://www.naa.gov.au/recordkeeping/er/web_records/

Cyberspace is a fluid environment in which communities evolve and die. It depends heavily upon complex interplays between identity, trust, and imagination. It is an environment in which information is given place and purpose. Information survives there as long as it is functional or until someone else presents it in a more exciting way.

So, in conclusion we must wonder what the cultural asset is that we must seek to manage: *Is it the content, the environment, or the experience*. While I would argue that it is all three, I recognise that even to preserve the content poses formidable challenges.

There are eight challenges we will need to meet if we aim to ensure the future has access to our internet culture. We need to:

- (1) develop local, national, and trans-national selection strategies that ensure that the by-products of our culture that are worthy of preservation are preserved and that their preservation proceeds in a coherent and structured way;
- (2) define what kinds of materials should be classified as 'worthy of preservation' and ensure that these definitions reflect the need to retain popular culture along side all other materials;
- (3) establish a typology of web content that will allow us to manage its harvesting, cataloguing, storage, and preservation more efficiently;
- (4) foster the creation of ethnographic studies of cyberspace communities;
- (5) establish a software archive because, no matter how much talk about XML (eXtensible Mark-up Language) offering us an independent and risk free preservation format, future providers of information will continually find new ways of presenting data and new kinds of data to represent. These will depend often on the use of proprietary software;
- (6) construct mechanisms to enable access to digital media that are independent of any current hardware and software—to ensure that these archive copies of harvests can be accessed in the future;
- (7) address the real challenge, which is to preserve the interactivity, immediacy, and reality of the net-based experience; and,
- (8) build collaborative initiatives that depend upon shared policies that are implemented in a layered manner.

Preservation can be approached from three vantages:

- ◆ a centralised model in which one group in one place tries to capture everything;
- ◆ a decentralised model where different groups capture material of interest to their institution, organisation, or nation; and
- ◆ a distributed model, where organisations collaborate on tackling preservation.

The distributed model where information creators (e.g. commercial organisations, government agencies) and national libraries, archives, and museums work together to document the intellectual and cultural heritage

that is being created on the Internet is the only viable solution. It leverages effort, ensures consistency of standards, and reduces redundancy of activity and duplication of content. The Internet is a collaborative environment and that is the only way we will be able to manage its preservation as a cultural asset and a social space.

Bibliography

Anderson, B., (1983), *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, (London: Verso).

Appadurai, A., (1996), *Modernity at Large: Cultural Dimensions of Globalization*, (Minneapolis: University of Minnesota Press).

Arms, W.Y., Adkins, R., Ammen, C., and Hayes, A., (2001), 'Collecting and Preserving the Web: The Minerva Prototype,' *RLG DigiNews*, 5.2.

Black, E., (2001), *IBM and the Holocaust: The Strategic Alliance between Nazi Germany and America's Most Powerful Corporation*, (New York: Crown Publishers).

Ceruzzi, P.E., (1999), *A History of Modern Computing*, (Cambridge, MA: The MIT Press).

Curtis, P., (1996), 'MUDDing: Social Phenomena in Text-based Virtual Realities', in Peter Ludlow (ed.), *High Noon On The Electronic Frontier: Conceptual Issues in Cyberspace*, (London: The MIT Press).

Dibble, J., (1998), 'A Rape in Cyberspace', in Richard Holeyton (ed.), *Composing Cyberspace: Identity, Community, and Knowledge in the Electronic Age*, (Boston: McGraw-Hill)

Dibbell, J., (1996), 'A Rape in Cyberspace; or How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database into A Society', in Peter Ludlow (ed.), *High Noon On The Electronic Frontier: Conceptual Issues in Cyberspace*, (London: The MIT Press).

Donath, J.S., (1999), "Identity and Deception in the Virtual Community" in M.A. Smith and P. Kollock (eds.) (1999), *Communities in Cyberspace*, (London: Routledge).

NARA Guidelines, (2001), *Guidelines To Agencies On Preserving A Snapshot Of Their Web Sites At The End Of The Clinton Administration*, (Washington DC: NARA).

Holeyton, R., (ed.), (1998), *Composing Cyberspace: Identity, Community, and Knowledge in the Electronic Age*, (Boston: McGraw- Hill).

Ito, M., (1996), 'Virtually embodied: the Reality of Fantasy in a multi-user Dungeon', in David Porter (ed.), *Internet Culture*, (London: Routledge).

Jones, C., (1995), 'What Goes into an Information Warehouse?' *Computer* (Aug), 84-5.

Kahle, B., (1997), 'Archiving the Internet', *Scientific American* (March),
http://www.archiving.org/sciam_article.html

Kuper, A., (1999), *Culture: The Anthropologists' Account*, (Cambridge, MA: Harvard University Press).

Malinowski, B., ([1922], 1950), *Argonauts of the Western Pacific*, (New York: E.P. Dutton).

Malinowski, B., (1954), *Magic, Science and Religion, and Other Essays*, (New York: Doubleday Anchor Books).

Matthews, W., (2001), 'Webmasters flinch at "snapshots"', *Federal Computer Week*.

Millard, W. B., (1996), 'I Flamed Freud: A Case Study in Teletextual Incendiarism', in David Porter (ed.), *Internet Culture*, (London: Routledge) .

NAA's Archiving Web Resources: *A Policy for Keeping Records of Web-based Activity in the Commonwealth Government*,
http://www.naa.gov.au/recordkeeping/er/web_records/intro.html

Nightingale, V., (1996), *Studying Audiences: The Shock of the Real*, (London: Routledge).

Relf, F.A., (1999), 'PANDORA – Towards a national collection of Australian electronic publications',
<http://www.nla.gov.au/nla/staffpaper/ashrelf1.html>

Reid, E. M., (1996), 'Communication and Community on Internet Relay Chat: Constructing Communities', in Peter Ludlow (ed.), *High Noon on the Electronic Frontier: Conceptual Issues in Cyberspace*, (London: The MIT Press).

Reid, E. M., (1999), "Hierarchy and Power: Social Control in Cyberspace", in Smith, M.A. and Kollock, P., (eds.), *Communities in Cyberspace*, (London: Routledge).

Ross, S., (2000), *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*, National Preservation Office (British Library), Occasional Publication.

Ross, S., (1998), 'The Expanding World of Electronic Information and the Past's Future', in Higgs, E, (ed.), *Historians and Electronic Artefacts*, (Oxford: Oxford University Press), 6-28.

Ross, S., (1995), 'Introduction: Networking and Humanities Scholarship', S. Kenna & S. Ross (eds.), *Networking in the Humanities*, (London: Bowker Saur), xi-xxiv.

Shore, B., (1996), *Culture in Mind: Cognition, Culture, and the Problem of Meaning*, (Oxford: Oxford University Press).

Smith, M.A. and Kollock, P., (eds.), (1999), *Communities in Cyberspace*, (London: Routledge).

Tepper, M., (1996), 'Usenet Communities and the Cultural Politics of Information', in David Porter (ed.), *Internet Culture*, (London: Routledge).

Turkle, S., (1997 rpt), *Life On the Screen*, (London).