10 Small Scale Academic Web Archiving: DACHS

Hanno E. Lecher
Leiden University
h.e.lecher@let.leidenuniv.nl

10.1 Why Small Scale Academic Archiving?

Considering the complexities of Web archiving and the demands on hard-
and software as well as on expertise and personnel, one wonders whether
such projects are only feasible for large scale institutions such as national
libraries, or whether smaller institutions such as museums, university depart-
ments and the like would also be able to perform the tasks required for a
Web archive with long-term perspective.

Even if the answer to this is yes, the question remains whether this is
necessary at all. One could think that the Internet Archive in combination
with the efforts of the increasing number of national libraries is already
covering many, if not most relevant Web resources. Does academic or
other small scale Web archiving make sense at all?

Let me begin with this second question. The Internet Archive has done
groundbreaking work as the first initiative attempting comprehensive archiv-
ing of Web resources. Its success in accomplishing this has been revolution-
ary, and it has laid the foundation on which many other projects have built
their work. Still, examining what the Internet Archive and other holistic pro-
jects\footnote{In contrast to topical archives, I consider holistic Web archives to try to capture
either the contents of the whole Internet, or at least certain domains thereof, using
automatic crawling routines.} can achieve it is easy to discover some limitations. Since their focus of
collection is very broad, they have to rely on robots for a large part of their
collecting activities, automatically grabbing as many Web pages as possible.
This kind of capturing is often very superficial, missing parts located further
down the tree, many pages being downloaded incompletely, and some file
types as well as the hidden Web being ignored altogether.

In addition, since harvesting is performed automatically and in irregular
intervals there can be no conscious selection of resources, and no possibility
to consider or detect important contents whose existence might be very short lived or difficult to detect. Of course, the Internet Archive and other large scale Web archiving projects do feature special collections, where much effort is spent on developing “deep” collections around a chosen number of topics. The number of these topics, however, is very limited, and it is obvious that many research projects will have to develop their own archives.

A further difficulty is the restricted accessibility to the contents of holistic Web archiving projects. Currently the exact (former) URL of a document or website has to be known to be able to retrieve the data. Usually it is not possible to search these archives using descriptive metadata or full-text indexing services. Even if a full text search option was available, the lack of conscious resource selection would only provide a similar amorphous result as a Web search today.

Looking at these limitations it becomes evident that the Internet Archive and others are neither archiving the Internet in its completeness nor provide ways of access suitable for many academic or other research purposes. Small scale academic archiving thus becomes an important need. But, returning to the question raised in the beginning is this feasible?

There are some issues to be approached differently or to be considered if the organization planning a Web archive is of smaller scope and has only very limited human and/or hardware resources. To illustrate some of the main issues, the Digital Archive for Chinese Studies (DACHS) will be examined as a case study. It should be kept in mind, though, that different projects will frequently need different approaches or solutions.

### 10.2 Digital Archive for Chinese Studies

The main objectives of the DACHS are to identify and archive Internet resources relevant for Chinese Studies in order to ensure their long-term accessibility. Selection plays an important role in this process, and special emphasis is put on social and political discourse as reflected by articulations on the Chinese Internet.

Currently the DACHS project is handled by the libraries of two Sinological institutes, namely, the Institute of Chinese Studies at Heidelberg University, Germany, and the Sinological Institute at Leiden University in the Netherlands. The organizational infrastructure is thus quite different from big national libraries (see Table 10.1).

---

2 Accessible at http://www.sino.uni-heidelberg.de/dachs/
Table 10.1. DACHS collection holdings (2005)

| What                | Number of files | Size in GB |
|---------------------|-----------------|------------|
| Discussion boards   | 2,04,349        | 1.6        |
| Documents           | 2,35,449        | 4.3        |
| Donations           | 979             | 0.15       |
| Films               | 982             | 0.427      |
| Journals & Newsletters | 2,62,939    | 5.8        |
| Websites            | 12,65,857       | 24         |
| Total               | 19,70,555       | 36.277     |

When the idea of downloading online resources from or about China was first brought up in late 1999 in Heidelberg, it was by no means clear what this would mean. The still much undeveloped idea was introduced as a possible part of larger application for creating a European Center for Digital Resources in Chinese Studies aiming at the improvement of the conditions for China-related research and information access in Europe. The project included a number of activities such as purchasing a wide range of commercial full-text databases, supporting the development of academic database projects as well as developing own ones, developing finding aids for printed and non-printed resources on China, and, above all, providing free access to all resources as widely as possible. The project was granted a term of five years, including the financial means to improve an already existing hardware environment and to hire some personnel in the form of student assistants.

The main guideline of the project was put as “maximum flexibility with maximum accountability”. This simply meant that there was room to develop detailed plans for many sub-projects on the go and even ideas for new sub-projects as needed.

While the year 2000 was spent to get most of the sub-projects in more advanced planning status running and to reorganize the IT infrastructure of the institute, concrete planning for the Web archive only began in 2001.

### 10.2.1 Initial Steps

Looking at the existing infrastructure from which the Web archive had to be developed, it became very obvious that the limits were tight indeed. The Digital Archive for Chinese Studies was to be run by the library of the

---

3 For a detailed description of the European Center for Digital Resources in Chinese Studies and its sub-projects please visit the project's homepage at http://www.chinaresource.org/
institute, which at the same time also oversaw the IT environment with its then four servers and close to 100 workstations. Responsible for maintaining this IT infrastructure was the librarian with the help of one or two part-time student assistants and – if needed – support from the ICT department of the University.

For the Web archiving project an additional part-time student assistant could be hired for handling the actual workflow (downloading, archiving, metadata creation, etc.), while the librarian – in addition to his responsibilities for the library and the IT environment – would take care of project management and conceptual development. In fact, the contribution of the project assistant to the conceptual development of DACHS was substantial. Involving the assistant to a very high degree was considered important to avoid concentration of knowledge about the theoretical framework of the project in the hands of only one person. This strategy paid off when the project management was seamlessly passed on after the librarian left the institute for another workplace.

Of course a number of issues had to be considered at the beginning of the project, taking into account the size of the institute and its possibilities. What would it mean to aim at long-term accessibility of archived resources? What were the requirements on hardware and software to create and maintain such an archive? How the selection of resources should be organized as an ongoing task, and how should the data be made accessible? And above of all: what else needed to be considered for proper planning, and where to look for answers?

Answers to the last question could be found in a document that in 2003 was to become the standard framework for Internet archiving, canonized as ISO 14721:2003, but generally known as OAIS, the Open Archival Information System. This document proved to be of crucial importance to our project since it provided us with much needed theoretical background and helped us to pinpoint most of the critical issues of Web archiving.

---

4 In 2001 the library of the Institute of Chinese Studies was home to some 1,10,000 volumes and staffed with one full-time librarian and about 20 part-time student assistants.
5 Two Sun Sparc stations running various online full text as well as bibliographic databases, a Linux machine serving our WWW pages and e-mail services, and a Novell server providing the fileserver infrastructure for the institute.
6 The librarian accepted the position as head librarian of the Institute of Chinese Studies at Leiden University. It is thus no accident that Leiden later became the first new project partner of DACHS.
7 The document is available at http://www.ccsds.org/documents/650x0b1.pdf. Cf. also Chap. 8 of this book.
Useful as this document is, it has one major drawback: as its name suggests it is a framework only, giving but very theoretical guidance on the various issues and leaving the concrete implementation to the ingenuity of the user. It became thus necessary to look for other places in order to come to an understanding of how to put the framework into practice. Much information could be found at places such as PADI, RLG and others, but even more important was the active participation in workshops and conferences that were dealing with issues related to Web archiving.

10.2.2 Institutional Sustainability

One of the most important questions the whole project had to face was how to provide institutional sustainability for such a long-term archiving project. None of the three major factors that were decisive in this question could be taken for granted on a long-term basis: financing of the project could run out; interests of people in the institute could shift, leading to negligence of the project; and even the long-term status of the institute itself was by no means guaranteed. It was thus obvious that an institute represented a far less trustable place for long-term archiving than a national library or archive, where legal provisions forced the institution to fulfil its responsibility towards the collection basically forever.

Strategies had thus to be developed to make the survival of the archive possible even after the institute itself ceased to exist or was not able anymore to support the project. Survival could be defined in two ways: either it should be possible to keep the archive active, which means that all activities from resource selection to ingest to making the data accessible would continue; or the archive should at least be preserved in a deep frozen status, which means that although no new data come into the archive, at least the accessibility of the already existing ones should be ensured.

The way to accomplish this is again twofold. First and foremost, the archive has to fulfil the basic attributes of a trusted repository, which means adherence to acknowledged standards as described in the OAIS model. This would make it possible for other institutions – most ideally a national library – to take over the contents should we not be able to accommodate the archive anymore.

The second way was to develop the archive into a distributed effort. If a number of institutions were actively and interactively participating in the project it would be possible for one partner to take over an archive from another institution that had to discontinue. Here again the adherence to

---

8 Cf. list of resources at the end of this chapter.
established standards were essential. For our project it was decided to
work towards a collaborative project and to look for possible partners as
soon as most of the local issues were solved and the archive had won recog-
nition as an essential contribution to the field of Chinese Studies.

10.2.3 Hardware

To ensure the proper working of DACHS on the local level a suitable
hardware environment had to be set up. Next to providing scaleable server
space and dedicated workstations for the download routines and manage-
ment purposes we had to give special consideration to backup facilities,
fall-out security and virus protection.

The computer center of Heidelberg University provides a sophisticated
backup system using the IBM ADSTAR Distributed Storage Manager
(ADSM). Using this system a backup of the whole Archive is made every
night onto magnetic tapes stored at the computer center. Regular backup
copies of these tapes are also stored at the University of Karlsruhe. The
data of the archive are thus kept at three different places which offer
reasonable security.

An Uninterruptible Power Supply (UPS) device as well as a RAID
(Redundant Array of Inexpensive/ Independent Disks) system (level one)
was installed to provide basic security for uninterrupted availability. For
virus protection, we turned to McAfee Virus Scan software. Using virus
definitions fetched on an hourly basis from the McAfee server, all incom-
ing data are routinely checked on viruses, and cron jobs automatically
incite regular scan processes of the whole archive.

Basically we could make use of an already existing IT infrastructure at
our institute which also included backing by the University ICT depart-
ment for the more intricate parts. Much of the above should be standard
equipment for universities anyway, mostly however, not within an institute
but rather maintained at central ICT facilities. This can be an advantage
since a more professional environment can be counted on. It may also be a
handicap, though, if restrictions on hard- and software apply.

10.2.4 Software

In order to actually start with the Digital Archive for Chinese Studies
project a piece of software was now needed that had to fulfil a number of
conditions.

The Internet can be seen as a huge collection of interlinked data in vari-
ous formats and encodings. Archiving such data means to preserve them in
a way that content, functionality and look and feel are kept as close to the
original as possible. Since look and feel of Web content does vary depend-
ing on the browsing software there is no way to preserve this in a reliable
way. However, it was felt that if we could preserve the original bit stream
most of the requirements above should be met. When downloading whole
websites, the original file structure should be kept intact. Of course, to
keep linking between the downloaded documents intact all links had to be
converted into relative links on the fly. We needed thus configurable
crawling software that was able to perform this task. The software should
be able to handle a wide range of different formats, including dynamically
generated contents of billboards and the like. And it should be affordable.
After some testing we found that Offline Explorer Pro by MetaProducts
provided all needed functionality, at least for the beginning.

10.2.5 Metadata

An issue that did engage us considerably was the creation of metadata. On
the one hand metadata offer an important access point for users since they
provide harmonized information describing the content of the document,
such as author, title, and subject. This kind of access is particularly impor-
tant as long as no full text search can be provided. At the same time, technical
metadata are unanimously described as essential to assist in long-term
preservation since they have to carry all sorts of information necessary for
proper administration and future handling.

Metadata are very costly to create, though, since even with semi-
automated metadata harvesting routines much of the work has to be done
by hand. Very obviously, it is impossible to create metadata for hundreds
of thousands of documents at the rate they are downloaded from the Inter-
net (see Fig. 10.1). And even if this were possible it proved to be very hard
to find out what exactly these metadata should contain, to what degree of
granularity, and in which format.

One of the questions that arose from these considerations was whether
or not metadata were necessary at all. Or better: would it be possible to
solely rely on search algorithms to retrieve all data needed for access as
well as for long-term preservation purposes? From the user’s perspective it
was argued that full text search was a much more reliable tool for finding
documents than crude subject headings or title words contained in often
imperfect metadata. From the technical point of view much of the neces-
sary information such as file format, download date, encoding, etc. should
easily be retrievable from the data themselves, or else could be made part
of the file naming structure.
In the end this idea was rejected. Digital resources are an extremely volatile form of information, and separate documentation on these resources is necessary to ensure their integrity and accessibility for the long-term future. Depending on the data themselves for this kind of information makes them extremely vulnerable and prone to rapid decay.9

To make the task manageable and to better integrate the archived material into the holdings of the library at large it was decided to create metadata as part of our regular catalogue. The catalogue was thus adapted to

---

9 Discussion and a framework of preservation metadata for digital resources can be found at the homepage of the joint OCLC/RLG project “PREMIS – PREservation Metadata: Implementation Strategies”, http://www.oclc.org/research/projects/pmwg/.
accommodate the necessary additional metadata, including categories for rights management, history of origin, management history, file types, identifiers, and others. Depending on the complexity of the resource we now create metadata records that may either describe a single file, as in the case of single text-only documents, or a whole set of files if the archived object is a website, discussion board or newspaper.

10.2.6 Collection Policy

Quite obviously the aim of the Digital Archive for Chinese Studies cannot be to preserve the Chinese Internet in its entirety. Neither is this technically an option, nor do we think this to be very useful. As a research institute, we are interested in parts of the Chinese Internet that reflect certain aspects of the Chinese society, and that are particularly ephemeral. Providing an informed selection of resources is thus an asset that helps current as well as future users to detect material we believe to be relevant. Of course values will change over time, and later generations might well have preferred different choices. In this case the Internet Archive still provides a very rich alternative where other choices will be available. It is the combination of the two approaches – selective and holistic – that provides the future user with the widest array of possibilities for his or her research.

At the same time, however, we do want to reduce the danger of being too narrow in our selection. As will be discussed below, DACHS is evolving into a larger cooperative project. Different partners will have different selection criteria and thus apply different sets of values to their selection process. This can only enrich the contents of the archive while the policy of informed resource selection remains intact.

Still, in order to make best use of the limited resources available to us we have to develop clever collection strategies. Given the number of Internet sites in or about China, and the pace in which relevant discussions begin, develop and disappear on the Chinese Net, the task of building a meaningful archive of selected online resources is a daunting one. To cope with this situation we started to build an information network of individuals (native or foreign scholars and “netizens”) who are actively or passively part of the discourse we try to grasp. Making use of their judgment and knowledge it becomes possible to identify specific sites or discussion processes that fit into our selection profile. The larger this information network the more diversified this part of the selection process can be.

Usually the information network also serves another purpose. Since its members are part of the contemporary Internet culture they can provide valuable context information about the archived resources. As much as
possible of this context information is created and preserved either in the metadata themselves or on dedicated Web pages that form part of the collection.

Since a few months we are also working on a number of special projects such as contemporary poetry, SARS, or the homosexual scene in China. Scholars and MA students working on these topics try to create comprehensive archives, including endangered, significant or representative websites, as well as other material such as photos and posters. To put these resources into context they add introductory texts as well as their own research material to the archive. We consider this kind of value added approach with its contemporary contextualization as essential for a better understanding of these volatile resources especially in the distant future.

But we do not solely rely on informants and scholars. Certain events of international impact such as the September 11 terror attack or the Olympic Games in China in 2008 frequently cause heated discussions on various platforms on the Internet. To capture such outbreaks of public opinion we are working through checklists of relevant discussion boards and newspapers, resulting in a set of snapshots of relevant material covering a time span of a few weeks before and after such events.

Somewhat similar to the special projects described above are collections donated or sold to DACHS by private persons, researchers, research groups, or other institutions. Sometimes, we are even approached by the publishers of endangered websites (or we contact them) to help them preserve the content. These collections where not specifically created for DACHS (and may thus not always follow all quality standards), but DACHS helps to keep these resources available on a long-term basis.

10.2.7 Partnerships

We have seen earlier that running the archive as a distributed collaborative effort involving coequal partners is an essential surviving strategy for small scale Web archives. Partnerships may help to ensure long-term preservation goals through offering inheritance service for members that have to terminate operations. They also allow for a distribution of workload, thus providing means of cost reduction and the possibility of a wider selection of archived resources. Hardware, experience and quality standards can also be meaningfully shared and might improve the overall performance of the archive. Not the least important are partnerships as part of a political

10 See e.g., the poetry section of the DACHS Leiden division at http://www.sino.uni-heidelberg.de/dachs/leiden/poetry/ as an example.
Recognizing these issues DACHS has begun to develop guidelines for possible partnerships. These guidelines are being explored together with Leiden University in the Netherlands, which is participating in the project since the end of 2003. The baseline assumption is that partners should be able to remain as independent as possible and keep their own identity, while there are a number of standards and services that need to be shared.

One of our major concerns is to build common finding aids, including a hyperlinked subject guide (or table of content) as well as a full text and a metadata search facility integrating the complete archive with all its current and future project partners. This not only requires a central access point in the form of a joint home page, but also considerations about how these search options need to be constructed to serve the desired purpose. Full text search must be possible across different domains, and for the metadata either a union catalogue is needed where all data are physically stored together, or a virtual catalogue has to be constructed that searches the various local metadata and presents them in a coherent manner. In any case there is a strong need to establish shared standards for the creation of the metadata.

Other issues that need to be discussed include a common access restriction policy, regulations for the division of labour, and routines that help to avoid duplication of efforts when archiving resources more than once.

Important for the design of the cooperation is also the degree to which prospective partners are able to either build a fully-fledged local archive infrastructure, or to make use of the facilities of one of the project partners. Since all the above issues are still work in progress it is impossible to share experiences at this point of time. It is well possible that more issues will emerge in the process of negotiation and actually implementing the partnership, but it has become quite clear that for small scale Web archiving projects cooperation on many levels is essential.

10.3 Lessons Learned: Summing Up

After having built DACHS for about four years now, many questions and problems still remain unresolved or in development. Looking back, however, there are a few issues we would try to deal with differently today than we did in the beginning.
Most essential in this respect is the allocation of positions for the project. In addition to the daily working routines of data ingest, the task of developing and managing a Web archiving project must not be underestimated. The decision to put this task exclusively into the hands of a librarian who is already performing a plethora of other duties is at least questionable. While his involvement into the development of the project has many advantages and important, a dedicated managerial position for DACHS would have been is much better suited to lead the project into smooth operation. Many issues could have been addressed much earlier and more effectively, thus positively effecting the project’s development.

A second issue that – at least from today’s perspective – should be solved differently is the choice of harvesting and archiving software. At the time DACHS started operation in 2001 Web archiving was still in its infancy, and many big players of today only just started to develop and publish their efforts. Although the software we chose for our purpose at the time fulfilled most of our needs to our satisfaction, the tools available by now are much better suited for the task. Without going into the details of possible candidates – this is done elsewhere in this book – it remains clear that a new choice will have to be made for our project in order to streamline the ingest process and creation of metadata.

The above may also illustrate a final point I would like to make. Many issues that make Web archiving difficult and in times even discouraging need not be solved by yourself. There is no need to worry about building a file format repository or developing your own harvesting software. It is possible – even necessary – to rely on the efforts of others. Collaboration does not only include partners of your archiving project, but also colleagues and institutions of the Web archiving community providing solutions and tools that you would not be able to create if on your own.

10.4 Useful Resources

The following list of resources is only a very small selection of what we found useful for our work. Many more resources will be found if you visit the websites below, especially the excellent subject gateway to digital preservation, PADI. All resources were accessible in May 2006.
10.4.1 Websites

Council on Library and Information Resources (CLIR)
  http://www.clir.org/
Electronic Resource Preservation and Access Network (erpaNet)
  http://www.erpanet.org/
International Internet Preservation Consortium
  http://netpreserve.org/
Internet Archive
  http://www.archive.org/
Networked European Deposit Library (NedLib)
  http://www.kb.nl/coop/ NedLib/
PADI Preserving Access to Digital Information (National Library of Australia)
  http://www.nla.gov.au/padi/
PADI: Web archiving
  http://www.nla.gov.au/padi/topics/92.html

10.4.2 Mailing Lists

Archivists
  http://groups.yahoo.com/group/archivists/
DigiCULT
  http://www.digicult.info/pages/subscribe.php
DIGLIB – Digital Libraries Research mailing list (IFLA)
  http://infoserv.inist.fr/wwsympa.fcgi/info/diglib/
OAIS Implementers (RLG)
  http://lists2.rlg.org/cgi-bin/lyris.pl?enter=oais-implementers
PadiForum (National Library of Australia)
  http://www.nla.gov.au/padi/forum/
Web-Archive
  http://listes.cru.fr/wws/info/web-archive/

10.4.3 Newsletters and Magazines

CLIR Issues
  http://www.clir.org/pubs/issues/
DigiCULT Newsletter
  http://www.digicult.info/pages/newsletter.php
D-Lib Magazine
  http://www.dlib.org/
DPC/PADI What is new in digital preservation
  http://www.nla.gov.au/padi/qdigest/
RLG DigiNews
  http://www.rlg.org/en/page.php?Page_ID=12081