"Digital Library - IT Opportunities and Challenges in the New Millennium"

Preserving current public discourse for the future: The Digital Archive for Chinese Studies (DACHS)

Jennifer Gross and Hanno Lecher, University of Heidelberg, Germany
Beijing, 9.-11.7.2002

Introduction

Internet as platform of public discourse

When you are browsing through journals and on-line discussion groups relevant for librarians, or when you are looking at the issues raised at conferences on library matters such as this one, you will easily notice that *"digital library"* is currently **the** dominating topic. However, within this topic the main interest focuses on digitization of print resources and their management, especially in the light of preservation and accessibility. Only recently another issue came up and is now increasingly discussed, namely capturing and preservation of material that has mainly been produced for on-line access, either through the World Wide Web, the Usenet, or other Internet services.

Compared with traditional media such as printed material, microforms, and even audio and video recordings, as well as digital material such as CD-ROMs, or digitized print resources, Internet resources are of a very different nature: for the first time no publisher serves as intermediary between author and readership. So, access to channels of publication suddenly is open to everyone, and the already overwhelming mass of commercial publications libraries have to cope with is now topped by a literally uncontrollable and unmanageable flood of daily changing contents. Traditional ways of handling library acquisitions are futile with on-line publications, so this new resource has so far been largely neglected and is heavily underrepresented in most library collections.

On the other hand, the fact that participation in the publication process has never before been so simple and effective for the common public, has two important implications: first, the Internet has developed into a platform of public discourse that is of increasing importance for society and politics, and not the least for scholars. The Internet will soon become one of the major resources for scholars of many different areas, to understand society and the twists and turns of public discourse.

Secondly, however, articulations of public opinion on the Internet are of a very elusive nature. The Internet is an ever changing kaleidoscope of contents, and although it is thus capable of representing development and diversity of social discourse very well, it is hard if not impossible to systematically keep track of what happens there. What is even more important, articulations on the Internet that have been made in the past are lost if we do not find ways to preserve these for the future.

An important project that tries to address this problem is the Internet Archive (http://www.archive.org/). Since October 1996 large parts of the global Internet are scanned every few months and stored for later research purposes. Useful as this may be as first aid measure, many problems remain unresolved: 1) as of today, full text or keyword search of the archive is not possible - you have to know the exact URL of a former Web site for access; 2) most of the Web sites are only captured very superficially, with parts located further down the tree not available, many pages being incomplete, and some file types being ignored altogether; 3) and harvesting is performed in irregular intervals,

without giving any consideration what so ever to important changes or articulations that have appeared in between.

For this reason many more initiatives using different approaches have come up recently. Some are of a more holistic character, such as projects started by various National Libraries that aim at preserving all on-line publications within their realm of responsibility. Many others work on a smaller scale, focusing on special topics, and giving much attention to appropriate selection criteria. What all these projects have in common is that they try to develop or follow standards for detailed metadata creation as well as consistency and quality issues. The most widely accepted of these standards is the reference model for an Open Archival Information System (OAIS), that is currently being reviewed as an ISO Draft.

Digital Archive for Chinese Studies (DACHS)

The *Digital Archive for Chinese Studies (DACHS)* is a project following this kind of approach. It is part of the *European Center for Digital Resources in Chinese Studies (ChinaResource.org)*, which was founded at the Institute of Chinese Studies at the University of Heidelberg in Germany.

Simply put, DACHS *"[...] aims at identifying, archiving and making accessible Internet resources relevant for Chinese Studies, with special emphasis on social and political discourse as reflected by articulations on the Chinese Internet"* (mission statement). Simple as this statement reads, a lot of questions arise from it: What does *archiving and making accessible* mean? What are *resources relevant for Chinese Studies*? And where is *social and political discourse* reflected on the Chinese Internet?

Collection Policy

Since the concept of national borders is alien to the Internet, articulations reflecting the Chinese social and political discourse may come from very different sources all over the world, including China proper, Hong Kong and Macau, Taiwan, Overseas Chinese communities, Chinese foreign students, as well as scholars, institutions, and mass media covering the Chinese speaking region. The term *"Chinese Internet"* is thus taken in a very broad sense, encompassing resources in all languages, and from all over the world. The archive will contain a broad range of different material, such as speeches from leading Chinese politicians, historical documents from American or Russian archives, non institutional Web sites created in China or elsewhere, clippings from Chinese discussion boards, and many others.

Identification of relevant resources

Given the limited institutional and financial resources available to us - after all, we ar not a national library, not even a University library - strategies of selection and cooperation are crucial for the success of the project. One way we do this is to first identify moments of heated debates on the Internet and then to clunch down on the relevant material that has appeared there. For this we make use of what I would like to call our "information network", that is the judgment and knowledge of individuals of all professions - foreign scholars and native Chinese - who frequently use the Internet and are (actively or passively) part of the discourse we try to grasp. This "human approach" implies a lot of deficiencies, to be sure, such as a significant portion of chance in identifying relevant resources, limitation to a very small fraction of the available resources, and a considerable amount of labor involved in the process of selecting, downloading, and metadata creation.

On the other hand we are thus able to very flexibly respond to current threads of discussion, we are able to consciously select a broad range of different opinions on various current affairs, and we can make full use of the background knowledge our informants provide, since that could be integrated as commentary into the set of metadata created for the resources.

Integration of external collections

In addition to resources gathered in this way by our own staff we also aim at extending our archive considerably by integrating complete collections donated or sold to the Institute by other parties (private persons, researchers, research groups, institutes or other organizations). These acquisitions will form special collections where different levels of access restrictions can be implemented, depending on the conditions under which they were given to us.

Legal issues

A major issue in the whole process is the question of copyright. There is an obvious cleavage between the necessity to archive resources of high significance for later research that would otherwise be irrevocably lost, and the wish to adhere to national and international copyright law. There has been much discussion on this topic, and the stances various governments have taken vary significantly.[1]

We believe that the following is a reasonable approach that trys not to infringe on current copyright law while at the same time - and this is important! - ensuring the future availability of resources that we think are of utmost significance for the academic community and the society in general.

As a general rule we will archive all resources we identify as being relevant and that are freely available on the Internet. Access to the documents and resources we have stored is restricted to password owners, and applicants must provide information on research purpose and institutional affiliation before being granted access. From within the Heidelberg University campus there is no password restriction.

However, should archiving be explicitly prohibited or should the copyright owner protest we will try to negotiate a solution that is acceptable for both parties, including payment of a royalty and/or implementation of complete or partial access restriction of the material in question. We already have designed the outlines of a more sophisticated access policy allowing easy implementation of various levels of restriction, which will become especially useful with the acquisition of external collections.

---

[1] Recently the European Parliament and the Council of the European Union have published a *"Directive on the harmonisation of certain aspects of copyright and related rights in the information society"* that aims at homogenizing the various legal approaches within the EU in this respect, and that is to be implemented by the member states by the end of 2002. Article 5 of the directive provides rules for exceptions and limitations to what is called *"reproduction right"*, i.e. the exclusive right of authors, performers, producers etc. to authorise or prohibit any form of reproducing their works.

These exceptions also include "[...] specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives which are not for direct or indirect economic or commercial advantage".

Working routines

Download routines

Now, how do we work?

Depending on the material we have developed three different approaches for getting hold of relevant resources:

First of all we try to single out certain "long term" topics such as China‘s relationship with the WTO, on which we are actively searching and collecting relevant material of all kind, making use of Internet search engines, newsgroups and mailing lists.

A second important focus are single events such as the September 11th terror attack or the NATO bombing of the Chinese embassy that cause heated discussions on the Internet. To capture such outbreaks of public opinion we are building up a check list of relevant discussion boards, newspapers, and Web sites, which will be worked through each time an important event happens. The result is a set of snapshots of relevant material covering a timespan of a few weeks before and after the event.

In addition to these two main approaches we also randomly collect fragments of public discourse that are believed by our researchers and informants to be of some relevance for current or later research and that neither belong to event related discussions nor pertain to one of our special collection topics.

Depending on these approaches and the kind of material we want to capture, we decide whether to apply regular downloads, irregular snapshots or single non-recurring downloads. Some categories such as single documents etc. clearly belong to non-recurring, complete downloads. On the other hand, discussion boards, some of them growing by hundreds or thousands of postings per day, can only be included in form of snapshots of a few week's discourse.

In the case of complete Web sites that we believe to be of major interest we will ensure automated download in regular intervals with additional downloads whenever we notice important changes or additions. In this we again depend on the help of our "information network".

Metadata creation

One of the most crucial and most time consuming parts of our working routine is the creation of metadata. On the one hand these metadata offer an important access point for users since they provide standardised information on author, title, subject, etc. On the other hand, in the case of digital resources and in view of their long term preservation metadata are of even higher significance since they have to carry all sorts of information on content as well as technical and administrative data necessary for proper identification and future handling.

For various reasons we have decided to put all metadata into one place, namely the library's catalogue. After consulting standards such as the OAIS reference model we have re-designed the catalogue to accommodate the necessary metadata, including categories for rights management, history of origin, management history, file types, identifiers, and others.

Depending on the complexity of the resource, metadata sets are created either for single files, such as in the case of single documents, or one record for whole Web sites, discussion boards or newspapers.

However, as the creation of detailed metadata is very time consuming and thus very expensive, the rapidly growing collection might call for different strategies and approaches to ensure accessibility and long term preservation. To solve this problem two approaches are being considered.

The first one is to use metadata harvesting routines. But since there is still a significant amount of "human labour" necessary to control and supplement the data, this approach might probably not be able to solve the problem.

A second solution could be to do without any metadata at all (or almost without metadata - of course there would be certain exceptions) and to try to rely on information that fulltext search engines can retrieve as well as on additional information that might be included into the URI of the object.

Access

Points of Access

There will be three options to access the collection: the project's homepage, the library's general OPAC, and a full text search engine.

Project's homepage

Currently the homepage of DACHS (http://www.sino.uni-heidelberg .de/archive/) provides access to its resources through a basic classification system, making use of certain keywords in the files' URL. These keywords are reflecting either nature (discussion boards, documents, films etc.) or topic (culture, economy, politics etc.) of a file. They are listed as life links on the homepage, delivering the corresponding records from the OPAC, from where direct access to the material is possible.

Catalogue

Another way to access the material is by simply using our library OPAC. ...

Full text search

Of course, one of the most natural ways to work one's way through digital resources would be the usage of a fulltext search engine. With the growing of the archive metadata alone will not be sufficient to help locating relevant material, so we have started recently to collect information on various fulltext retrieval systems that are able to index documents in different encodings (GB, Big5, Unicode etc.) and file formats (html, MS Word, and pdf, to name only a few).

Technical infrastructure

I won't talk too long about our computer system now, but of course we are fully aware that a well designed IT infrastructure is essential if you want to be successful in running something like a digital archive, something that aims at long term preservation of digital data.

Security issues

- Dedicated and climatized IT-room

- UPS (uninterruptable power supply -- protects server park from power supply problems)
- Software raid (level 1) (data are stored simultaneously on two harddisks)
- Daily ADSM backup to University Computer Center
- Additional backup to University of Karlsruhe
- Virus scan on download
- Daily virus scan of the complete archive
- Hourly update of virus definitions

Server

The server hosting the data of the archive is running on the Debian distribution of Linuxall the data is a Intel Pentium 3 machine (coppermine) with 700 MHz CPU, 60 GB of raid level 1 harddrive space and 256 MB RAM, runing on Linux Debian 3.0. The data are stored as a separate part of our Apache Web server that is connected to the Internet through a 100 mBit/s line.

Our complete IT equipment running the various servers and including switch and hub is installed in a dedicated and climatized room. UPS (Uninterruptable Power Supply)

To provide a certain degree of availability we have installed a software raid level 1. This system is based on free linux drivers compiled in the servers kernel 2.4.4 instead of special hardware components. It writes all incoming data onto two different harddrives, so the first one is a 100% copy of the second.

In addition to this we have also implemented a backup strategy using the IBM ADSTAR Distributed Storage Manager® (ADSM). Every night a backup of the whole archive is made onto magnetic tape at our University Computer Center. For additional security regular backup copies of these tapes are also stored at the University of Karlsruhe, some fifty kilometers from Heidelberg. Thus there are four copies of the archive allocated to different places.

The McAffee Virus Scan v4.14.0 for Linux is used to protect the collection. Cron jobs automatically incite regular scan processes of the archive. Infected files are re-moved to a save location and the administrator of the archive is given notice via E-mail. Every hour a perl-program checks the McAffee homepage for an updated version of the virus file.

Workstations

Two Workstations are dedicated to download and management purposes. One is for regular downloads and more or less self-operating. The other is used by the staff to search for new sites, establish best practices and options for regular downloads as well as to do all non-recurring downloads. Further more it will be used for cataloguing and administrative work.

Both computers will be running on Microsoft Windows 2000 NT. For the download process we either use the Microsoft Internet Explorer, if the object consists of one single page, or the MetaProducts Offline Explorer Pro 2.1 for complete Web sites or larger parts thereof.

On both download computers a local virus scan program is installed. By opening a file the program will check it for virus’.

Current status of the project

Done

Since the start of the project in August 2001:

- Begin of download activity from the beginning
- We have a small network of informants
- We have established a suitable IT infrastructure
- We have done some work on our metadata set

After six months of work our collection so far (June 2002) contains about 230.000 files, roughly corresponding to 2.9 GB in size.

What	Number of files	Size in MB
Discussion boards	184,853	1,400
Documents	1,378	59
Films	550	430
Full text databases	12,300	315
Journals & Newsletters	15,020	339
Monographs	21	7
Web sites	20,200	386
Total	**234,322**	**2,935**

To do

Improvement / fine tuning of metadata set

Implementation of search engine

Contact to other projects

Establishment of team of informants

Who is DACHS

Four people are part time working on DACHS responsible for different tasks:

1. The director (Prof. R. Wagner, the head of the Institute) is responsible for supervision, collection policy, and financial management. He is one of our busyest informants...
2. The content manager (Hanno Lecher, the librarian) is responsible for information about general developments of digital libraries and archives, recurring evaluation and development of concept, contents, contacts as well as organization and control of the other staff, and finally supervision of the archives integrity.
3. The assistant content manager (Jennifer Gross, a student worker) is responsible for download and archiving, control of access points, conceptual development and supervision of the archives integrity.