Learning by Doing: The Digital Archive for Chinese Studies (DACHS)[1]

Jennifer Gross

University of Heidelberg, Germany

Abstract. This paper introduces the Digital Archive for Chinese Studies project. This comparably small project is aimed 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. The paper outlines the history, the collection policy, technical infrastructure, legal position, working routines, points of access, as well as the current status and future steps of the project.

An Idea was born

In recent years research in Chinese Studies underwent a fundamental change. In former times scientists were accustomed to study China by doing either field research or reading primary and secondary sources in form of printed material stored in libraries and archives. The first method is rather expensive terms of time and money, the second can never provide up-to-date information, but lags behind at least the time it takes to send something from China to the country in question.

This situation changed with the invention and expansion of the World Wide Web. By going on-line researchers can dive into Chinese public discourse: read the news, participate or monitor discussion boards, search for articles, books, music and films, or else surf on the waves of the Chinese Internet. Hence, the Chinese World Wide Web became an important resource for scientists in at least two aspects: it provides up-to-date information on recent developments and is in itself a field to be studied as it reflects a lively public discourse, a discourse that was rarely accessible to scholars of Chinese Studies as it is unfiltered by publishers and often uncensored by any state, especially Communist China.

The Internet is an ever changing kaleidoscope of contents, thus being able to represent the development and diversity of social discourse. However, its elusive nature makes

---

[1] Parts of this paper have been presented at the „Digital Library - IT Opportunities and Challenges in the New Millenium" conference, Beijing, 8-12 Jul 2002 by Hanno Lecher and Jennifer Gross.

it hard if not impossible for scholars to systematically keep track of what happens. And even more important: the past and present of the Internet will be lost forever, if nobody takes care to preserve them for the future.

As the significance of these developments became more and more obvious, the head of our Institute for Chinese Studies, Prof. R. Wagner, gave birth to the idea of a depository for Web-born material that might be of interest to Chinese Studies now and in the future. The idea was further developed by our former librarian Hanno Lecher and found its place in between the holdings of our more conventional library – consisting of monographs, periodicals, video and audio recordings as well as databases – as the *Digital Archive for Chinese Studies*. It is part of the *European Center for Digital Resources in Chinese Studies (ChinaResource.org)*, which was founded in 2001 at the Institute of Chinese Studies at the University of Heidelberg in Germany.

Launching the Digital Archive for Chinese Studies (DACHS)

At the start of the project only one thing was absolutely clear: we had to start downloading as soon as possible.

Our goal was: *"[...] 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).

What this really meant and how it could be achieved, we found out on our way. It started with the definition of a collection policy and development of a technical infrastructure. Then we had to clarify our legal position and establish best practices for working routines. And finally we had to make the collection accessible to the public by building a Web interface. Now let me describe in detail how far we got, the failures we experienced and what remains to be done.

Collection policy

What is the "Chinese Internet" ?

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. For example, the archive contains the selected works of chairman Mao, an on-line published regional Chinese newspaper, clippings of discussion

boards after the September 11th bombing, historical documents from Russian archives or American non government Web sites of China watchers.

Identification of relevant resources

Given the limited institutional and financial resources available to us - after all, we are not a national library, not even a University library - strategies of selection and cooperation have been and still are crucial for the success of the project. Basically we rely on the knowledge of what I would like to call our "information network": 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. Some belong to our institution's staff, but most are scattered all over the world communicating, of course, via the Internet.

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 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 is integrated as commentary into the set of metadata created for the resources.

Integration of external collections

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

Technical infrastructure

As DACHS was not our Institutes first project focusing on digital resources we could rely on a well designed IT infrastructure and an experienced IT team right from the start of this project. But of course, to develop and host a digital archive providing long term storage and access to digital data is a task not to be taken lightly.

Server

Our main server hosting all the data is a Intel Pentium 3 machine (copper mine) with 700 MHz CPU, 60 GB of raid level 1 hard drive space and 256 MB RAM, running on Linux Debian 3.0. The data is stored as a separate part of our Apache Web server that is connected to the Internet through a 100 MBit/s line.

The McAffee Virus Scan v4.14.0 for Linux (updated automatically every hour) 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.

Workstations

Two Workstations are dedicated to download and management purposes. Both computers are 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 viruses.

Backup strategy

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 hard drives, so the first one is a 100% copy of the second.

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

Legal position

Right from the start of the project a major issue was 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.

We believe that the following is a reasonable approach that tries 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 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.

Working routines

Download

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 these active search for and collection of relevant material is done, making use of Internet search engines, newsgroups and mailing lists.

A second important focus are single events that cause heated discussions on the Internet or for which relevant information often cannot be found elsewhere. Examples are the 16th congress of the Chinese Communist Party or the outbreak of SARS (Severe Acute Respiratory Syndrome) in Asia. 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 period 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 “information network” to be of some relevance for current or later research and that neither belong to event related discussions nor pertain to a special collection topic.

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 weeks’ 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.

Further more, important journals are included in the collection and downloaded according to their publishing frequency.

Metadata creation

One of the most crucial and most time-consuming parts of our working routine is the creation of metadata. These metadata offer an important access point for users since they provide standardized information on author, title, subject, etc. Moreover, 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 Information Model[2] we have re-designed the catalogue[3] 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 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 full text search engines can retrieve as well as on additional information that might be included into the URI of the object.

---

[2] "Preservation Metadata and the OAIS Information Model. A Metadata Framework to Support the Preservation of Digital Objects." A Report by the OCLC/RLG Working Group on Preservation Metadata: http://www.oclc.org/research/pmwg/pm_framework.pdf, June 2002.

[3] The library's catalogue adheres to the AACR2 rules as defined in "The Concise AACR2. 1998 Revision. Michael Gorman. Chicago: American Library Association, 1999."

Points of access

Currently there are two options to access the collection: the project's homepage and the library's general OPAC. We are still looking for a full text search engine that can handle the diversity of file types and encodings as well as the mass of documents our collection already includes.

Project's homepage

Currently the homepage of DACHS (http://www.sino.uni-heidelberg.de/dachs/) provides access to its resources through a basic classification system, making use of certain keywords in the files' URL. These keywords reflect 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, though, only with login and password.

Catalogue

Another way to access the material is by simply using our library OPAC. The OPAC is accessible via LAN from within the institute or Internet. It provides simple or combined search methods for author or creator, title or title words, and subject headings, among others.

Current status of the project

Done

Since the start of the project in August 2001:

- Start of download activity from the beginning
- Establishment of an “information network”
- Establishment of a suitable IT infrastructure
- Development of a metadata set
- Development of best practices on different sorts of download tasks
- Collection of about 630.000 files, roughly corresponding to 9 GB in size

To do

- Further improvement / fine tuning of the metadata set according to the needs of the collection

- Further establishment of the “information network”
- Testing and implementation of a metadata harvester
- Testing and implementation of a search engine
- Promoting the project

Conclusion

The Digital Archive for Chinese Studies project shows how a small institution with modest financial and personal resources – three people are actively working on DACHS: a supervisor and two part time student workers doing the downloads and metadata creation – can build, manage and maintain a highly specialized Web archive. Most important to the success of the project has been the selection process by our “information network”. This enables us on the one hand to reduce the flood of information on the Web to a maintainable and preservable amount, on the other hand, we can be sure that only such information is included into the collection that is relevant from a specialist’s point of view.