Everything is not Lost
The Digital Archive for Chinese Studies (DACHS)

Simply put, the Digital Archive for Chinese Studies '[...] 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?

By Jennifer Gross & Hanno E. Lecher

Articulations reflecting the Chinese social and political discourse appear in all languages and, as the concept of national borders is alien to the internet, from all over the world, including China proper, Hong Kong and Macau, and Taiwan. The Digital Archive for Chinese Studies (DACHS) defines 'Chinese internet' in a very broad sense, to include material from overseas Chinese communities, Chinese foreign students and from scholars, institutions, and mass media covering the Chinese speaking region. It is the aim to cover a broad range of resources on varied topics, such as the reflection of the SARS epidemic on the internet, speeches from high-ranked Chinese politicians, historical documents from American or Russian archives, non-institutional websites created in China or elsewhere, and clippings from Chinese discussion boards.

A human approach

Strategies of selection are crucial for the success of the project. After detecting outbursts of heated internet debates as much as possible of it has to be captured . To this end we rely on our 'information network', that is the judgement and knowledge of frequent internet users from all academic fields, various professions, and every nationality, who are (actively or passively) part of the discourse concerned. This human approach harbours many deficiencies, to be sure. Not only does the selection, downloading, and metadata creation process require a considerable amount of labour, identifying relevant resources in this fashion also involves a significant portion of chance, and in the end only a tiny fraction of the available resources can be covered.

The name of this Chinese website is 'Don't cry, my friend' (*pengyou bie ku*). The site hosts a nation-wide 'Dating & Personals' service for gay and lesbian Chinese. Although the site tries to operate within the boundaries of national law, it might very well be closed down one day. The DACHS project can help to make sure that its contents will be still available when that day comes.

Having said that, this approach will enable us to respond to current threads of discussion, to consciously select a broad range of different opinions on various current affairs, and to make full use of the background knowledge our informants provide, due to the fact that we will integrate the latter as commentary into the metadata for the resources. In addition to gathering resources in this fashion, we also aim to considerably extend our archive by integrating complete collections donated or sold to the Institute by private individuals, researchers, research groups, institutes, and other organizations.

As there is quite some variation in our focal material, we have developed three different approaches for getting hold of relevant resources. First of all we try to single out certain broad, long-term developments such as China's relationship with the WTO and the like. On these topics, we are actively searching and collecting a wide range of relevant material, making use of internet search engines, newsgroups, and mailing lists.

A second important focus are historical events that cause heated discussions on the internet, of which the debates concerning the 11 September terrorist attack are an obvious example. To capture as much of such outbreaks of public discussion as possible, we are setting up a list of relevant discussion boards, newspapers, and websites. These are checked whenever an important event occurs, so as to compile a set of snapshots, covering a time-span of a few weeks before and after this event. Third, certain fragments of public discourse may neither pertain to event related discussions nor to one of our special collection topics yet be considered of some relevance for current or later research. Essentially, the process of collecting this miscellaneous material, is slightly

A shared technical effort

In order to provide the best possible access to the material it is necessary to create metadata. This process ranks among the most crucial and most time consuming parts of our working routine. For one these metadata offer an important means for access as they provide standardized information on author, title, subject, and so forth. Moreover, in view

of the long-term preservation of digital resources metadata are highly significant in that they can carry all sorts of information on content as well as on technical and administrative data necessary for proper identification and future handling. At present, the library catalogue of the Institute of Chinese Studies in Heidelberg has been re-designed so as to accommodate the necessary metadata, including categories for rights management, history of origin, management history, file types, identifiers, and others.

The DACHS has been developed largely according to the Open Archival Information System (OAIS) framework, an ISO standard for long-term preservation of digital data. In the DACHS issues such as daily backup routines in three places in southern Germany, failure-proof IT systems, and virus checks based on hourly updated virus definitions are cared for in the best way possible.

Obviously, a comparatively small institute such as ours cannot satisfactorily cover the field on its own; hence preparations for international cooperation are now in progress. While it is necessary to distribute the effort, it is essential that the integrity of the whole is preserved. Together with the Institute of Chinese Studies at Leiden University, where currently large digital collections on the SARS epidemic, contemporary poetry, and the homosexual scene in China are being built up, DACHS is currently running the first testing phase for transnational cooperation. In close cooperation with the Leiden University Library possibilities of large digital resource management systems are also explored and will most probably be used for the whole project. ‹

Jennifer Gross (University of Heidelberg) is manager of the DACHS project at the Institute of Chinese Studies, Heidelberg University, and currently writing an MA thesis on press control in the People's Republic of China.
jenny@gw.sino.uni-heidelberg.de

Mag. phil. Hanno E. Lecher is librarian of the Institute of Chinese Studies, Leiden University. His current research activities include social networking among Buddhist gentry of the late Ming dynasty, as well as aspects of modern library infrastructure development.
h.e.lecher@let.leidenuniv.nl

Preserving Endangered Sites

Due to its worldwide proliferation over the past few years the internet has become an important scholarly resource in at least two aspects: it provides up-to-date information on recent developments, and it is a field to be studied in its own right as it reflects a lively public discourse unfiltered by traditional publishing processes and sometimes even uncensored by the state.

However, articulations of public discourse 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, 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. 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 important problems remain unresolved: (1) most of the websites 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; (2) scans are performed at irregular intervals, without considering possible important changes or articulations that have appeared in between.

For this reason many other initiatives using different approaches have come up over the past few years. Some of these projects, such as those started by various National Libraries are of a more holistic character and aim to preserve all online publications within their realm of responsibility. Others work on a smaller scale, focusing on special topics, and paying much attention to appropriate selection criteria. Currently maintained at the Institute of Chinese Studies in Heidelberg and with the Sinological Institute in Leiden as a new partner, the *Digital Archive for Chinese Studies* (DACHS) follows this approach in capturing articulations on the Chinese internet. ‹