Nov.11, 2004

Harvesting the Web, Preserving Chinese Voices.

The Digital Archive for Chinese Studies (DACHS) in Heidelberg

Rudolf G. Wagner,
Professor of Chinese Studies,
Director, Center for Digital Resources in Chinese Studies

with Jennifer Gross, DACHS content manager,

University of Heidelberg, Germany

The Internet[1] has become an important medium of communication in the Chinese-speaking and the sinological world. It has empowered many people who hitherto had no public voice to join in public debates, has allowed public voices to venture into domains formerly out of their reach, and has become an important archive for sources of relevance for Chinese Studies now and in the future. Given the communication structure of the PRC with a strong and unified censorial control by the CCP, many public utterances concerned with state and society will only be accessible

[1] The DACHS content manager MS Jennifer Gross has been responsible for the technical information on DACHS in this article. She and Mr. Nicolai Volland have been kind enough to join me in a discussion to map the outlines for this article. They have compiled and analyzed the data contained in the three Appendixes. I also have benefited from critical notes by Mr. Michael Luedke, Dr. Thomas Kampen, and Prof. Barbara Mittler.

for a very short time (this is true even for official government statements), and many of the more elaborate statements or even works from within the PRC will only surface in the public Internet domain outside the PRC. Even then, their availability over the longer term is precarious.

There has been some optimism that the economic opening of China would lead to a relaxing of the censorial controls. This optimism has not been borne out by actual developments, and even after the Jiang Zemin/Hu Jintao transition the massive efforts to control public articulation have been continued, and have been publicly advertised.[2]

A Memory Preserved

---

[2] Just a short note from a November 2004 article will highlight this:

Six Political Commentators Blacklisted to BeExpurgated from Media

Shortly after Mr. Wang Guangze, the reporter and commentator of theThe 21st Century Business Herald, being dismissed from hisposition, Beijing has tightened its grips onfreedom of speech, according to the Hong Kong based Apple Daily.The CCP's Propaganda Bureau issued an order about three weeks ago to theofficial media to expurgate articles by and reports about Jiao Guobiao andother five political commentators from the press.

Informed source in Beijing says Jiao Guobiao, associate professor of Beijing University, has written articles appealingBeijing togive people the right of election and to abolish the Propaganda Bureau.Jiao is now in the United State for a short visit, deliveringspeeches to advocate his political opinion.

Other blacklistedcommentators include Li Rui, senior member of the CCP, who keepsappealing the authorities for essential political reform, Wang Yi and YuJie, writers who usually write on hot political topics, and Mao Yushi,director of Tianze Economic Research Institute headquartered in Beijing, who writesarticles critical of the CCP's economic policy. Many feel queer to see Yao Lifa was blacklisted for expurgation too, because being an activist indefense of the right of peasants in Qianjiang, Hubei Province, Yao does not write so much as the others.His addition to the blacklist, it is alleged, may be promoted by the CCP's worry over the mounting problems in the rural area.

The source also indicates that scholars and journalist inChina has feltthe tightening control of speech since Hu came to power in lastSeptember and many have come to the conclusion that Hu is inferior toJiang in open-mindedness.

An anonymous scholar in Shanghai says that Hu's refusal to reevaluatethe 1989 Tiananmen Massacre and his arrest of the New York Timesjournalist Zhao Yan indicate he is more reckless than rational in dealingwith critical affairs, so there may be less hope than expected in China's future political reform."

In order to preserve these voices, and make them accessible to China-scholars, the Digital Archive for Chinese Studies has been set up in the framework of the European Center for Digital Resources in Chinese Studies, Heidelberg University, in 2001. The roughly 1.5 Mio Chinese-language sources stored in DACHS (about 80% of the total) are supplemented by a daily inclusion of relevant documents, newspaper articles and TV broadcasts in Western languages, ranging from new releases of the Cold War Archives to various Newsletters, from statements and documentations by institutions such as the US Department of Defense, Amnesty International, or religious associations to articles in the press. Already now, many studies on Chinese public reactions to particular events benefit greatly from this resource which is beyond the reach of a memory administration by the party/state routinely and often retroactively excising parts of the historical record.

At the same time, the increasing use of WWW resources in scholarly articles highlights the instability of this medium. A short investigation will serve to highlight the highly problematic nature of using Internet sources as evidence in scholarly writing. For this purpose, a random sample of three recently published articles has been examined. The three articles, two from *The China Quarterly* in 2003 and one from *China Information* in 2004, rely to a significant degree on Internet resources for their main arguments. This means that they do not only quote web-based sources as supporting evidence, but rather use these sources to build their argumentation. As a consequence, their hypotheses, and thus the verifiability of their arguments, depend directly on these resources remaining available to reviewers, readers, and other scholars.

The investigation results are as follows:

Article	Number of Internet sources used	Still available as of Oct. 13, 2004	No longer available
1	10	5	5
2	4	4	0
3	8	6	2

Only the Internet sources used in article 2 are still accessible as of October 2004; half of those quoted in article 1 are no longer available, and 25 per cent of those in article 3. None of the articles that have disappeared from the web (usually the browser shows

an error message) could be retrieved by other means, such as search engines or the famous Google cache. These sources – if they have not been archived anywhere else – are as a matter of fact lost, and lost is thus the possibility to check on the articles in question.

Some further remarks are necessary: articles 1 and 2 quote sources from inside the PRC, while article 3 draws on websites both from China and from abroad. Two of the three foreign websites quoted in article 3 are no longer available. The problem is thus not simply one of the PRC controlled domains, but of a more general one. Of the PRC websites, articles hosted by large government-owned websites such as those of the People's Daily and the New China News Agency tend to remain available after several years; the same is the case with some local administrative bodies. The situation looks most bleak with local newspapers, but also with government bodies such as the People's Bank of China. However, even the resources that remain must be considered fragile, as changes in political direction might very well lead to the deletion of sources openly available today.

Finally, a most problematic factor concerns time. All three articles have been published in 2003 and 2004, the research for them was done only little earlier. The electronic sources for the three articles in question were last accessed by the authors between late 2001 and early 2003. In the short time span of 2-3 years, a significant amount of crucial sources has thus disappeared and is no longer within reach of academic researchers. None of the articles, furthermore, has used the most fragile kind of material in the first case, dissident websites or non-governmental material relating to sensitive issues that might easily fall prey to turns in a political tide. The majority of sources used are unobjectionable to the present Chinese government. How difficult the situation is can indeed be demonstrated best by a poem Prof. Michel Hockx from the School of Oriental and African Studies in London posted on the Chinese web in the course of research for his article on the literary website Rongshu.com: in his last footnote Hockx assumes that sooner or later his creation might be deleted by the website editors, declining to speculate on how long this might take. By the time the article appeared, the poem had been lost in the depth of cyberspace. Details on all articles in Appendix 1.

In a second round we will conduct a random check through the data collected by DACHS during the past 3 years. The files have been thematically grouped into 74 blocks, for each block one random file was checked. The result is less disturbing: 66%

of the articles can still be found at the original URL, 18% more can be found via the title and Google, 15% are gone. Details in Apendix 2.

In a third round, we will look at the life-expectancy of URL in the field of Chinese-language WWW discussion. One of the rather dissident websites , 思想评论 Sixiang pinglun , appends a list of links to URL with similar contents, the **友情链接**. Of the 59 items on this list from 14. Nov. 2001 , three years later only 34 were still alive. 4 of them could be found with a direct hit; three more were automatically redirected; 8 were indirect hits through small corrections of the original URL, and no less than 19 were indirect hits through google searches of website's name. Any loss here signals loss of an entire trove of information, not just a single document. Details in Appendix 3. The development in Chinese Studies mirrors that in other fields of research.[3]

My presentation will focus on harvesting strategies (and difficulties), indexing the files, search engine selection as well as cooperation in developing this resource.

The Digital Archive for Chinese Studies (DACHS)

In early summer 2001, the Institute of Chinese Studies in Heidelberg started to develop the Digital Archive for Chinese Studies, DACHS. DACHS is part of the European Center for Digital Resources in Chinese Studies set up in Heidelberg with the generous support from the Alfried Krupp von Bohlen und Halbach Foundation. Today, DACHS contains 1.515.877 files, corresponding to 25.2 GB. Of these 169.420 files of 3.1 GB are China-related individual digital source materials, scholarly, newspaper as well as web-based articles in different languages (mostly Chinese and English); 229.198 files of 4.5 GB are regular automatized full downloads of relevant

---

[3] DPC/PADI "What's new in digital preservation" write: "Following their influential study published in Science in 2004 (vol. 302, pp. 787-788), Dellavalle and colleagues have produced more evidence of the volatility of Web references in the medical literature. Their recent studies have included references in oncology journals (Hester, et al., 2004) and the Internet citation policies of high-impact STM journals (Schilling, et al.,2004). These support other recent studies of URL references in the the biomedical literature (Crichlow, Davies & Winbush, 2004; Wren, 2004); the stability of Web references has also been cited as being a critical issue for the publication of clinical trials (Tumber and Dickersin, 2004, pp. 278-279). Bar-Ilan and Peritz (2004) have provided a similar analysis of Web documents in the informetrics sub-discipline of information science." (http://www.dpconline.org/graphics/whatsnew/issue9.html#2

Chinese and Western language information resources on China. The large rest consists of full and updated downloads of BBS and web-sites. As DACHS contains large amounts of copyrighted material, it is accessible to scholars free of charge with a password individually provided by the Institute upon application or for whole research institutes through their IP range.

Reasons for setting up DACHS

There is consensus among people in charge of collecting information resources and keeping them available that the World Wide Web is a rich, a diverse, a valuable, a messy, and an unstable information resource. It does not only offer digitized information that is otherwise available in printed form, but has rapidly become a medium with a very large amount of original information not existing in any other medium. The disappearance of this type of information from the WWW means that it is permanently lost. Such loss is not intrinsically bad. Mankind would instantly drown if even the totality of the information on this world alone for a single minute would be stored in its totality. A total downloading and permanent storage of all information available on the WWW would not do more good than if a library would buy all books published in the world. However, while there might be differences of opinion concerning selection criteria, the notion that some selection has to be done to keep data from redundancy, seems ot be generally accepted. As a consequence, many institutions devoted to preserve and keep accessible information have felt prompted to start devising ways to select, store and make retrievable these data. They face the difficulty of developing a meaningful harvesting strategy.

A look back at earlier phases of library development shows the problems. There is a routine rule for librarians all over the world not to collect schoolbooks, and not to collect tract literature. The consequence has been that two excessively important sources for the development of modern education, of a “national” consciousness of the citizens of a given state, and of value changes, are not to be found in the major collections of the world. The sudden surge of interest in such sources suggests that resource collection strategies have not only the burden to satisfy the actual users of scholarly libraries, but anticipate the needs of future users as well.

The data themselves are of different sorts. Some data are designed for long-term use such as full-text databases, others come in ephemeral media. With the shift

in scholarship interests away from the grand personalities and the signal events to the historical experiences and cultural fabric of the common people including their preoccupations and tastes, scholars now want access to materials deemed unworthy of collection a short while ago. The large body of ephemeral data on the WWW is in itself part of this broader cultural sphere, and contains a large amount of the data future scholars are going to study in their research.

The Chinese situation is particular in various respects and I will focus on those most relevant for the development of DACHS.

- The data are voluminous, multilingual and multiscript. Alone in the Chinese characters, at least Big5, Guobiao and Unicode have to be accommodated, and reasonably also Japanese and Korean scripts including their own codes for Chinese characters. Other data come on HTML, PDF and other formats. This leads to considerable problems for search engines offering a unified access portal.
- The data might theoretically also exist in a printed form, but due to the diffuseness of the PRC Chinese print market and the insufficient funds of Western libraries to buy and store even a large amount of the Chinese print sources, the digital form becomes the factually only practicable access to these data abroad and even in China itself.
- As everywhere, a very large amount of the available information repeats other data. Here, we already see a particularity of the PRC media. They are under the unified control of the Central Propaganda Department of the CCP. For certain areas such as foreign relations or leading CCP personalities, the New China News Agency retains the publishing monopoly. It alone is allowed to circulate articles on these subjects which then have to be carried or summarized by the PRC papers. There is a regular stream of directives from the Propaganda Department as well as its local branches banning certain topics and prescribing the way in which others have to be written. The seeming variety of, for example, newspaper publishing in the PRC thus contrasts with an actual dearth of opinion and information. It does not seem meaningful to download this material into DACHS. The only meaningful use of these data would be for studies of the PRC media control. For this purpose, however, the

large number of digital archives maintained by PRC papers themselves are quite sufficient.

- At the same time, particular political constellations time and again provide temporary but official shelter for types of information and opinion at the margins of the prescribed discourse. A well-known example is the *Nanfang zhoumo*. While always moving within the prescribed limits of officially sanctioned campaigns against corruption and the like, it tested the limits. The limits were defined, for example, by the rank of the Party cadres involved in corruption who could be named, by the need to show the exceptional character of corruption cases, and by the need to show that there was no broad social acceptance of corruption among Party cadres and as a consequence no above-board corruption. The reports in this paper together with reports on criminal cases provide important information on the difficulties of the transition to a "socialist market economy." The relevant data, however, are stored in the server of the paper itself. Over the last couple of years, time and again key personnel of the paper have been removed, and with them each time a relevant section of the incriminated articles disappeared from the paper's server. Papers closed down altogether for infringement of the Propaganda Department's rules will have their URL closed so that access to the digital version of the paper is gone altogether. This retroactive censorship eliminates relevant scholarly information. The latest example is just a few days old. On Nov. 1, 2004, just before the US election, the *China Daily* ran an article by former Foreign Minister Qian Qichen, "US strategy seriously flawed," that was highly critical of the Bush administration. Its Chinese version had appeared on Oct. 18, 2004 in an open Party paper, *Xuexi Shibao*, which is published by the Central Party School. Even before Bush's victory, controversy arose in Peking about this article and various Chinese sources indicated to the BBC and other new agencies that its publication in the China Daily hat not been vetted by the Propaganda Department. Whatever this information is worth, and it certainly is utterly implausible, the article was promptly and retroactively removed from the *China Daily* website. This is where an entity such as DACHS can be useful by establishing at least the outlines of a memory of Chinese society outside the control of the Propaganda Department of the CCP.

- Given the strict hierarchization of information access in the PRC as well as the maintenance of a stratified neibu/gongkai divide, and the concomitant importance of getting inside information given the powers of the Peking Center, an inordinately large amount of information and opinion about all aspects of the PRC is not available through official channels. It circulates by word of mouth, but might also enter the higher levels of the public sphere such as foreign newspapers or scholarly reports. Journalists and scholars working and publishing beyond the borders of the PRC play a crucial role in making this information accessible to the scholarly world. Needless to say, this information, like any other information, might contain, besides facts, falsehoods, misunderstandings, well-guarded painful secrets, simple slander and occasionally even things legitimately classified as state secrets. In a communication structure strictly controlled as that of the PRC, this informal flow of information becomes a key medium of the public's following the political process. Whether we have to do with popular ditties satirizing government practices or personnel, SMS messages on the danger of SARS at a time when the government denied the existence of this threat, inside stories of meetings of high-ranking officials and the factional line-ups or the complaints of a Shandong farmer about supplementary fees leveled by the local government, this broad stream of information is of relevance for the study of this huge society and therefore DACHS has made substantial efforts to archive it and keep it accessible. The information might turn up in an article by an Australian journalist, in a Hong Kong Chinese-language paper, in a more systematic background article in a journal or even in a scholarly article.
- Up to the present the Chinese government, the CCP and the Chinese military have only made a very modest and reluctant use of the WWW to disseminate their own information. In fact, due to the control over the official media, one can reasonably assume that these institutions might reasonably assume that their official information and opinion is reproduced there. Accordingly, there is no perceived need for separate web-sites. The Propaganda Department does not have a web-site, and neither does the CCP or the PLA. At the same time, there are institutions within the central government and the regions which have made public on their WWW-sites large amounts of especially social and

economic data. Examples would be the Statistical Bureau, the Ministry of Agriculture, and recently the Central Audit Office that publishes selected cases of official corruption and the fight against it on its website. The reliability of these data is not without challenge, and so is the stability of their availability on the net. There is a good chance that they will be unavailable after a while, or retroactively changed to accommodate new priorities, assessments or information. They are, however, important historical information, independent of whether they are seen as propaganda, fact or the "factual" information available to decision makers at a given time. As many relevant data are hidden in the deep net and will be found only after some time-consuming digging by specialists, DACHS has downloaded them wherever feasible and wherever help was found to access and evaluate them. It would be as nonsensical to dismiss this official information as irrelevant propaganda as it would be to discount the informal information as just rumor and hearsay. While both charges might often be true, they circulate in the public sphere, reflect attitudes, have an impact, and might contain factual information.

- Finally, on a rapidly developing medium like the Internet, new spaces of societal interaction emerge that come to the attention of the party-state only with a time lag. In the interval between their emergence and potential intervention from the Propaganda Department, semi-autonomous niches come into existence that allow for non-official discourse to develop; these niches sometimes expand at stunning speed and involve significant numbers of a tech-savvy, well-informed online populace that is keen to explore the discursive limits of these novel spaces of interaction. If their potential to undermine the CCP's efforts to shape public opinion is noted, interventions may occur, which as a rule are followed by the disappearance of the debates from the web. DACHS makes efforts to stay abreast of such developments and preserve at least some of these sources. Examples for such autonomous niches are Weblogs (blogs), or online diaries run by individual users to communicate their ideas and interests with like-minded surfers – (crackdown started in summer 2004) – and FLASH videos that are increasingly used to propagate cartoons poking fun at official politics, but also for nonconformist political ideas, such as hyper-nationalism – (crackdown still to occur).

•

The WWW and Chinese public articulation

The WWW has become an important forum for Chinese public articulation. This is a strongly contested field. For a very short moment of a few months in 1999 and early 2000, the Chinese internet experienced its “golden age”. Government controls were not in place. Very quickly, however, a concept was developed to undo the very essence of the Internet, its openness and disregard for national borders. To maintain control over the public discourse in whatever form, the Center heavily invested in technology to establish a “Golden Shield” around the PRC that would prevent an easy and uncontrolled flow of information. The technology was supplemented by the establishment of a special Internet Police with the duty to filter out messages deemed morally or politically questionable and to selectively prosecute transgressions. The Shanghai Internet Police alone is said to count 1200 officers. Their main duty is the censorship of officially sanctioned chatrooms and bulletin boards, and the perusal of e-mail and sms messages picked up by screening technology focused on a regularly updated set of key words. In the Chinese-language domain a substantial number of sites have their location outside the PRC. They are often inaccessible from the PRC but there is a constant hide-and-seek between censors and users, and in fact people with the technical skills and the time required can often bypass official blocks, quite apart from the fact that shareware software is now available from the US that greatly facilitates such bypasses. In broad strokes, we thus have in the PRC itself:

- The huge volume of communications of SMS messages, mobile phones and telephones. These have shown, for example in the SARS case, to be important ways of citizen-to-citizen communication beyond government control. There are claims that the mobile phone has become an organizing tool for local protests. Government efforts have been stepped up to control these media and highly publicized arrests have spread the message that even this private communication is supposed to stay within government-prescribed bounds. Hardly any of these data are preserved, and their life-expectancy is real-time zero, accessibility for an entity such as DACHS is utterly marginal and dependent on individuals sending such data via WWW.

- Bulletin boards on official servers such as the 强国论坛 *Qiangguo luntan*. What appears there has been approved by censors but still is often of interest. These messages are inaccessible after a very short while. Their life-expectancy is counted in minutes, hours, or days. Real-time accessibility is given, but presupposes an enormous and unrealistic time investment to continuously download.
- E-mail communication among private persons. This communication is transnational in character and carries large amounts of documents which essentially are public communications, be they literary works, political essays, or scholarly articles. As there are only a few (namely four or five) routers linking the PRC to the WWW, the international traffic in this field is rigidly controlled. Life-expectancy of these data is short, accessibility is random.
- Private, but officially sanctioned Web-sites in many different fields, ranging from works of poetry and fiction to scholarly information such as the Bamboosilk site in Peking or an occasional think tank. These have a higher level of stability, but more often than not depend on the work input and ongoing enthusiasm of an individual to maintain it and keep it accessible and on the swings in policies enacted by the Internet Police. Their life-expectancy is counted in months and years, but steeply decreases after 3 or four years. Accessibility is good, but often unreliable.
- Data from official institutions and media. Apart from retroactive screening of often the most interesting material, these data do not, as a rule, have a much longer life-expectancy than some of the privately-maintained web-sites. Accessibility is good.
- Newly emerging forms of public communication, such as the FLASH videos noted above. Distribution happens through individual communication or dedicated websites. As they are working in uncontested spaces their long-term availability cannot be estimated but their status must be considered to be precarious.

There are many participants in Chinese public discourse who are aware of the CCP notion that defines the Chinese public sphere as being coterminous with the state

borders, and they disagree. Outside the PRC there is a substantial number of individuals and groups, and even commercialized entities who have made it their goal to keep Chinese data publicly – and that ideally also means inside the PRC - accessible without submitting them to the controls of the Peking center. They range from advocacy groups of a political, ethnic or religious character to scholarly networks and forums for open and public debates on what is seen as pressing issues of China. The authors are very frequently situated in the PRC and have their opinion published in this manner. Often a message initially circulating in the PRC via e-mail such as the essay proposing the abolishment of the Propaganda Department, "taofa Zhongxuanbu," by Prof. Jiao Guobiao from Peking University, will make its way abroad, and will then be made publicly available on web-sites situated beyond the PRC borders.

As these web-sites are mostly managed by individuals, they have one great advantage and two problems. Their advantage is that the people in charge follow certain sectors very closely, and have inside information about interesting new websites or new addresses for old ones which have been taken off the Web. The problems are that they are often agenda-driven in their selection, and their long-term availability presents a problem already now. A regular harvesting of these web-sites for securing long-term access and prevention of tampering is a must for an entity such as DACHS.

A number of Western institutions have begun to develop digital archives for particular sectors, among them the Carter Center for issues of democracy or the Berkeley China Internet Project center for issues of the Chinese public media and internet, or the British Library for the Silk Road. These institutions are stable in their structure and it is to be assumed that their data would be stored and kept publicly available if they should discontinue their operations in this field. To fully download their harvest into DACHS would not make sense, and a link, optimally allowing unified access to their holdings as well as those of DACHS through a single portal, would be best.

At the same time a very large number of individuals, mostly students, scholars, and journalists, have amassed highly focused sets of digital resources for their particular studies or work. These are very valuable, because they tend to be the result of knowledgeable selection not of random harvesting. Most of them are never made publicly accessible even though the people who selected them have finished their

work with them and would not object to their public availability. In many cases they include reference elements such as full indexes of a given periodical under study.

Technical Infrastructure and Format

As DACHS was not the Institute's 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 was not an easy task, and it is far from finished.

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 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, which will check each new outside file when it is opened.

Our main server hosting all the data is an 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.

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

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.

To provide a certain degree of constant 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 Heidelberg University Computing Center. For additional security, regular backup copies of these tapes are in turn stored at the University of Karlsruhe, some fifty kilometers from Heidelberg. Thus there are four copies of the archive allocated to different places.

Copyright

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 in conformity with EU copyright concepts. It tries not to infringe on current copyright law while at the same time - and this is important! - ensuring the future availability of important resources.

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.

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 standardized 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 Information Model[4] 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 expensive, the rapidly growing DACHS 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 labor" necessary to control and supplement the data, this approach might probably not be able to solve the problem.

A second solution could be to make 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.

---

[4] "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.

Cataloguing

As there is at the moment no full-text search engine installed, users have to rely, beyond author/title searches on subject headings. In order to anonymize the organization of DACHS, and to link up with search routines familiar to many scholars, the subject headings of the Library of Congress are used. The same is true for the library of the Institute, which also uses LOC shelf marks and MARC cataloguing formats. In this manner, the different printed, digital and image sources can be accessed through a unified set of headings.

Collection focus of DACHS

Is there and should there be a focus in the DACHS harvesting? Given the sheer volume, the very uneven quality, and the overwhelming levels of duplication there evidently is a need to select. DACHS is part of an institute with the goal of making accessible for researchers the widest possible array of research tools, and sources of scholarship in the entire field of Chinese Studies. It has a hybrid collection which includes source collections, scholarly monographs, periodicals, full-text databases of sources as well as of scholarship, films and musical scores. As a consequence, there is no criterion of content for DACHS (such as contemporary politics, Sichuan archaeology, or historical grammar) that would fit this agenda. At the same time, the overwhelming number of ephemeral digital data are in fact concerned with China today. This is even true for many subjects seemingly far away from the present such as DNA tracing or archaeology. The particular situation of the PRC public sphere outlined above thus creates a particular criterion for relevancy, namely items which have attracted public attention and have led to controversy in the Chinese-language world as well as China's interaction with the world. This is the realm where relevant data are most easily lost and where an independent archive will be most useful.

While this gives a rough criterion, it is quite evident that the actual decision making process is in the day-by-day perusal of accessible sources keeping in mind the actual and potential uses of the document in question. In fact, a "fuzzy" selection criterion that adjusts to the intrinsic potential of the individual item has proven to be the best. Instead of developing an abstract set of possibly inoperative criteria, let me

outline some of my decision-making processes in the last couple of months as the person who developed the concept for DACHS, and still is the main feeder.

On a newslist, Jiao Guobiao's article mentioned above appears. Evidently, this will be downloaded. A search is then made for more information about him, his publications, and his fate after this article. This nets an interview with Jiao in the Guardian on the background of this article and a few notes on Chinese web-sites. All of these will be downloaded. They become part of a large category of entries entitled "PRC media control".

A Canton paper publishes a list of the "fifty most important" public intellectuals in China. Perhaps the criteria are shaky, and the process of defining who should be in and who not is open to debate. Still, this might be an indicator that these people at least have some standing although none of them is in the Party leadership. The list is downloaded into DACHS. A search is initiated to find all available articles by and on these intellectuals on the web, and insert them into DACHS.

From a private communication on another matter, I hear that "Nature" published a "Letter to Nature" from a group of PRC authors in September 2004. It offers results of testing Y-chromosome markers of southern Chinese populations. As opposed to the known mitochondrial DNA tracing data published by Cavalli-Sforza, which showed that the southern Chinese population is genetically closer to the Tai-population than to the northern Chinese, who in their turn are genetically closer to the Mongols, the Y-chromosome data from the male populations show a much closer south-north link. Given the sensitivity of the topic to "national unity" and the efforts by PRC scientists to prove the genetic homogeneity of the Han-Chinese nation, the *Nature* article is downloaded together with other scholarly articles dealing with the historical process of interaction between north China and Northeast and Central Asia as well as with North-South Chinese interactions. Copies of Chinese articles (in *Acta Genetica Sinica*) about the genetic cohesion of the Han-Chinese are inserted into the offprint archive of the Institute.

Members of a newslist are alerted by a participant that, due to the Freedom of Information Act in the US, a document has been released by the US National Archives on a conversation between the Soviet ambassador and the US Undersecretary of State in 1969 in which the Soviet Union explored potential US reactions to a plan to use nuclear arms on the Chinese missile testing site in Xinjiang. This information is carried on the web-site of the Cold War Project and is likely to

remain available there. Still, because of its importance, and the clearer focus of DACHS on China-related matter, it is downloaded.

An article in a Western climatological journal is brought to my attention in which the results of the analysis of a stalactite found in a cave near Peking are reported. Due to the incremental and regular growth of the stalactite it provides rather clear historical sequence for climate shifts in China with the result that the north up to the twelfth century was in fact "green" and only then became cooler and, as it was reduced to a single rainy period, started drying out. This article would not easily come to the attention of sinological historians and therefore is downloaded.

In a footnote to an article, a young scholar refers to the author-title index she has compiled for a Chinese literary periodical from the late thirties she has studied. The periodical is in the holdings of the library. She is contacted with the offer to make this database accessible on-line via DACHS. She accepted.

At the present stage, DACHS offers a combination of full downloads of selected publications, data, news reports and opinion statements of interest and relevance for contemporary Chinese developments, and scholarly articles. It already contains numerous documents which for one reason or the other are not available elsewhere on the WWW, and provides targeted access to a wide variety of materials of high interest and quality, but located at places rarely accessed by Chinese Studies scholars.

Development plans for DACHS

DACHS at the moment is still strongly dependent on the input from a very small number of persons. In order to enhance its quantity without neglecting questions of quality, to enhance its visibility and usefulness, and to broaden the pool of qualified persons involved in harvesting, the following steps are being taken:

- Many scholarly journals in Chinese Studies publish articles which in their footnotes refer to on-line documents. A random check showed that within two years, many of the URL given in the notes have been closed and there often is no other place on the WWW where the document can be located. We have written to the editors with the offer to have their authors send the documents to

us for insertion into DACHS, from where they can be retrieved by readers with a simple password.

- Many scholars have amassed data from the WWW concerning a particular topic of their interest. These data become irrelevant for them after their work has been finished. We published an appeal through various specialized lists offering to insert these collections into DACHS with a tag of their name as the collector. If relevant and feasible, efforts will be made to continue collecting documents in the respective field.
- Cooperation with other participants is actively being pursued. At the moment, a cooperation with Mr. Lecher from the Institute of Chinese Studies, Leiden University, has begun. Mr. Lecher has moved to Leiden from Heidelberg, where he had been, among other things, in charge of the technical management of DACHS. At this stage, Leiden has found local scholars willing to input materials in the areas of Chinese contemporary poetry, SARS, and Chinese gay culture. Discussions are under way with another institution in Paris. On the technical cooperation side, DACHS is part of a project (in cooperation with the Prussian State Library in Berlin, the Institut fuer Asienkunde in Hamburg and the University Library in Goettingen) to develop and install a unified portal for information retrieval in East Asian studies. The project is funded by the German Research Council.
- The harvesting of the Chinese-language bulletin boards still is very insufficient. The same is true for targeted searches on subject-matters or individuals. At the present stage of development and planning, it will be meaningful to make efforts to find funding for the manpower needed to do and professionalize this important and very demanding work.

Problems of DACHS

At this stage, DACHS downloads cannot be retrieved with full-text searches, but only through keywords and/or author-title searches. The former are taken from the LOC keyword catalogue to secure an internationally accepted roster. In cataloguing entries we follow the AACR2 standard. Efforts to find a search engine capable to

handle the volume and diversity of DACHS traffic have not now been successful because search engines are either beyond our financial reach, or not powerful enough. Given the exponential growth of DACHS, this problem requires an urgent solution.

While a university institution such as the Heidelberg Institute offer a higher life-expectancy for the availability of the data collected, such institutions in fact are rather volatile. On the technical side, the storing and securing of the harvest requires a constant and high-quality professional management as well as costly hard- and software, for both of which money has to be found. On the leadership side, a labor-intensive entity such as DACHS requires a long-term commitment of an institution rather than just of a small number of mortals. Much work needs to be done in creating and maintaining a stable environment for such an enterprise.

Altogether, the Heidelberg Institute is happy that its foray with DACHS has met with considerable interest within Chinese Studies, but also in the wider community of institutions interested in WWW harvesting strategies. As will be seen from this presentation, at the very core of this collection is an intentional fuzziness that allows a very high degree of flexibility and innovation. We intend to keep it this way. At the same time, search strategies will have to be made more transparent so as to allow for other scholars as well as professional staff to become involved in the selection process itself.

Appendix 1: Life-expectancy of web-sites quoted in selected scholarly articles.

1. Linda Chelan Li. “The Prelude to Government Reform in China? The Big Sale in Shunde” in *China Information* XVIII (March 2004).

Sources last visited ca. Oct. 2002

10 WWW sources

location: PRC

result: 5 lost, 5 still accessible

details:

www.pbc.gov.cn/search_Result.asp X

www.news.online.sh.cn/news/gb/content/2002-03/10/content_309220.htm X

www.molss.gov.cn/index_tongji.htm O

http://202.130.245.40/chinese/PI-c/250186.htm O

www.cet.com.cn/20020718/yaowen/200207181.htm X

www.sc168.com/allnews1/topnews/200211010158.htm X

www.people.com.cn/GB/paper66/7740/737847.html O

www.sc168.com/titnews/shizhengz/2002shiliuda/..%5C..%5Cjinji..%5Csdzz..%5Csdaaa..%5C200210250622.htm O

www.conghua.gov.cn/chzf/zf001.htm X

www.sc168.com/outlook/shundehistory/200201100121.htm O

2. Joseph Fewsmith. “The Sixteenth National Party Congress: The Succession that Didn’t Happen” in *The China Quarterly* 173 (March 2003).

Sources last visited: unknown (before March 2003)

4 WWW sources

location: PRC

result: 0 lost, 4 still accessible

Details:

www.people.com.cn/GB/shizheng/252/8956/8960/20021022/847410.html O

http://news.xinhuanet.com/newscenter/2002-10/22/content_604652.htm O

http://news.xinhuanet.com/newscenter/2002-10/24/content_607017.htm O

http://news.xinhuanet.com/newscenter/2002-11/17/content_632239.htm O

3. Benjamin Penny. “The Life and Times of Li Hongzhi: Falun Gong and Religious Biography” in *The China Quarterly* 175 (September 2003).

Sources last visited: summer 2001.

8 WWW sources

Location: PRC and abroad

Result: 2 lost, 6 stil accessible

Details:

http://english.peopledaily.com.cn/special/fagong/1999072200A101.html O

http://english.peopledaily.com.cn/special/fagong/1999072200A102.html O

http://english.peopledaily.com.cn/special/fagong/1999072200A103.html O

http://english.peopledaily.com.cn/special/fagong/1999072200A105.html O

http://english.peopledaily.com.cn/special/fagong/1999072200A106.html O

www.cesnur.org/2001/falun_march03.htm O

http://209.196.48.36/photo/images/master_li_pics/heying/images/Master_family01_big.jgp X

http://209.196.48.36/photo/images/master_li_pics/heying1.htm X

4. Michel Hockx. “Links with the Past: Mainland China’s Online Literary Communities and their Antecedents” in *Journal of Contemporary China* 13 (2004).

1 WWW source.

Result: Lost.

Detail:

http://www.rongshu.com/rss/viewart.rs?aid=1236097

Appendix 2

Nr.	Titel	Still at old address	Found via Google	Not found
1	Chinese Media Suddenly Focus on a Growing AIDS Problem			1
2	″仁″字臆斷 ： 從出土文獻看仁字古文和仁愛思想 / 龐朴		1	
3	Export growth and the exchange rate : China in the middle of the East Asian Crisis		1	
4	China regulator warns of investment bubble /		1	
5	Reshuffle planned for newspaper group			1
6	我要揭露希望工程腐敗案 ： 我為什麼要離開青基會(5)			1
7	Court to deal with Internet crimes		1	

8	Dynamics of ethnic cultures across national boundaries in Southwestern China and mainland Southeast Asia : relations, societies, and languages	1		
9	Accountability without democracy : the principal officials accountability system in Hong Kong	1		
10	中国对片面追求GDP增长说“不”	1		
11	ROC (Taiwan) students in the U.S.A. (1950-2000)		1	
12	三峽庫區移民值得重視的幾個問題	1		
13	Reminiscences About the Reindeer Herders of China	1		
14	中國大饑荒 ： 一九五九至一九六一			1
15	Meaning, production, consumption : the history and reality of television drama in China	1		
16	China faces financial crisis at the turn of the century /		1	
18	Unbreakable spirits : women breaking down barriers in China	1		
19	No support for Taiwan independence, Nixon assured China in 1972 : new documents reveal origins of current U.S. policy	1		
20	China's Failing Health System	1		
21	Revisiting the Silk Road	1		
22	Taking academic games seriously	1		
23	論某公 ： 何新退出江湖的最后一篇政論			1
24	China makes a swift retreat on transparency			1
25	New Journalism	1		
26	Outline of classical Chinese grammar			1
27	China Has Tightened Genetics Regulation - Rules Ban Human Cloning;	1		
28	当代思想史上的“读书奖”事件	1		
29	五四運動與美國對於中國宣傳活動再論	1		
30	Glossary of chemical and biological warfare terms	1		
31	National identity in Taiwan after the lifting of martial law : a taiwanese nation in the making	1		
32	The Jiang proteges and the Jiang theory		1	
33	國家憲政体制的若干猜想 ： 一兩個局外人的對談錄			1
34	Conceptual metaphor theory as methodology for comparative religion		1	

35	"中國農民調查"觸及現實之痛(圖)		1	
36	PRC biotech : top researcher sees great prospects		1	
37	Half of new graduates in China can't find jobs : last year, 65% had jobs; the Sars outbreak and more longer-term structural problems are blamed		1	
38	我國公布首次自殺雕查結果 ： 每年28万人死于自殺	1		
39	The War on Terrorism: China's Opportunities and Dilemmas	1		
40	丁子霖女士给各国学者的一封信	1		
41	Ten defects in China's urban development, hilarity for things big and foreign	1		
42	Hu performs without song and dance	1		
43	Französischer Kotau vor Hu Jintao : Chinas Partei- und Staatschef in Paris	1		
44	WirtschaftsWoche kooperiert mit Chinas größter Wirtschaftszeitung		1	
45	China warns HK: Don't stray too far	1		
46	Asia keeps New Europe on edge	1		
47	US, India forge defence, security alliance to contain China		1	
48	Bad press' causing headaches for Beijing : Chinese leadership struggles to control freedom of expression	1		
49	Israel cancels exhibit in China		1	
50	Crisis in Tuscany's Chinatown	1		
51	China's Confidence in its "New Thinking on Sino-Japanese Relations	1		
52	Kiribati fears Beijing's new strategy	1		
53	Agreed framework between the United States of America and the Democratic People's Republic of Korea	1		
54	The erosion of one-party rule : clientelism, portfolio diversification and electoral strategy	1		
55	Aufgaben der deutschen Außenpolitik	1		
56	How to subvert democracy: Montesinos in Peru	1		
57	Clean elections and the "great unwashed" : electoral reform and class divide in the Philippines			1
58	A new "cult of personality" : Suslov's secret report on Mao, Khrushchev, and Sino-Soviet tensions, December 1959			1
59	Singapore in 2000 : continuing stability and renewed prosperity amid regional disarray			1
	China slams US report on Taiwan	1		

60	Chiang Rai looks a bit like China	1	
61	Tibet, its ownership and human rights situation	1	
62	Richest' lists grow in China : Briton uses public documents to name wary entrepreneurs; then Forbes jumps in.		1
63	新疆追記	1	
64	Mugabe hires China to farm seized land : half of confiscated plots are not being worked	1	
65	Declaration on the conduct of parties in the South China Sea	1	
66	四中全會及出台的″決定″意義重大	1	
67	Entering the Gate of No-Life	1	
68	Economy, DPP government gain from legislative score after special session	1	
69	US-CHINA materials are only sent to subscribers	1	
70	Full text of resolution on amendment to CPC constitution		1
71	National People's Congress : popularity and power		1
72	China's new cultural revolutionaries: they wave not Mao's book but eviction orders	1	
73	Breakdown : how China's decentralized health-care system is failing hundreds of millions, as diseases like SARS spread	1	

Appendix 3

Check on friendly pages linked by 思想評論 (http://www.sino.uni-heidelberg.de/archive/websites/sinoliberal011114/www.sinoliberal.com/), downloaded 14.11. 2001. The list is given at the bottom of the home page of the sixiang pinglun.

Nr.	Name	URL	Hit 1	Hit 2	Hit 3
1	公法评论	http://www.gongfa.com/default.htm	-	http://www.gongfa.com	
2	谢泳居	http://xieyong.yesky.net/default.htm	-	-	http://lookin.51.net/xy/
3	问题与主义	http://www.wtyzy.net/default.htm	-	-	-
4	思想论坛	http://www.sixiangbb	-	-	http://www.

		s.com/default.htm			16167.com/xilan/
5	制度分析与公共政策	http://www.wiapp.com/default.htm	-	http://www.wiapp.com	
6	世纪中国	http://www.csdn618.com.cn/century/index.htm	-	-	http://www.cc.org.cn/newcc/index.php
7	李慎之研究	http://lsz.tongtu.net/default.htm	-	-	-
8	中国哲学网	http://www.ehawk.org/default.htm	-	-	-
9	思想在线	http://www.thonline.net/default.htm	-	-	-
10	自由思想者网刊	http://gdkk.yeah.net/default.htm	-	-	-
11	政治文化研究网	http://www.tszz.com/default.htm	-	-	-
12	法律思想网	http://mylaw.myrice.com/default.htm	-	-	http://law-thinker.com/
13	行政法论坛	http://adminlaw.126.com/default.htm	-	-	http://adminlaw.netsh.net/
14	《万科周刊》经济人俱乐部论坛	http://www1.vankeweekly.com/asp/bbs2/list.asp@boardid=3	-	http://www1.vankeweekly.com/asp/bbs2/list.asp?BoardId=3	

15	笑蜀文集	http://www.zhenxiang.net/default.htm	-	-	http://www.boxun.com/hero/xiaoshu.shtml
16	读书网站	http://www.dushu.net/default.htm	-	http://www.dushu.net/	
17	学术的境界	http://oyjyy.tongtu.net/default.htm	-	-	-
18	墨闲居	http://moxianju.yeah.net/default.htm	Rd: http://go.6to23.com/moxianju/		
19	天涯之声	http://www.tianya.com.cn/cgi-bin/default.asp	-	http://www.tianya.com.cn/ - 天涯虚擬社區	-
20	秋天的乌托邦	http://fatwangyi.home.chinaren.com/default.htm	-	-	-
21	外国文学论坛	http://wgwx.tongtu.net/default.htm	+		
22	中国报道周刊	http://www.mlcool.com/default.htm	-	Rd: http://mlcool.edula.com	
23	中国学术城	http://xueshu.newyouth.beida-online.com/default.htm	-	http://xueshu.newyouth.beida-online.com/	
24	思想格式化	http://www.pen123.net/default.htm	-	Pen123.net	-
25	当当人文社科	http://www.dangdang.com/dd2001/makebr	+		

		w/01.03.00.00.00.00.asp			
26	胡星斗个人主页	http://huxingdou.home.chinaren.com/default.htm	-	-	http://www.univillage.org/huxd.htm
27	朝圣山之思	http://pilgrims.yeah.net/default.htm	-	-	http://pkuthinker.51.net/
28	学者庄园	http://go2.163.com/glz	-	-	-
29	任不寐个人主页	http://bmzy.126.com/default.htm	-	-	-
30	梅花网	http://www.mafaa.com/default.htm	-	-	-
31	公民教育网	http://www.citizenedu.org/default.htm	-	-	http://hongling.org/citizen/index.php
32	羊子的思想家园	http://lib.126.com/default.htm	-	-	http://yangzi.00books.com/
33	宪政文本	http://www.libertas2000.net/default.htm	-	-	-
34	北望亭	http://bwt.home.sohu.com/default.htm	-	-	-
35	中国经济在线	http://www.ceol.8u8.com/default.htm	-	-	-
36	观点周刊	http://go.163.com/~viewsweekly/untitled.htm	-	-	-
37	天虎评论	http://www.tyfo.com/tanfoplan/talk/default.htm	-	-	http://www.haodx.com/url/1857.ht

					m
38	思想的境界	http://www.100yearchina.org/www/main.html	-	-	http://www.open-society.com/sixiang/
39	小雅斋	http://go.163.com/~bluefire/default.htm	-	-	-
40	文化中国	http://www.culchina.com/default.htm	-	-	http://home.seechina.com.cn/seechina.html
41	素心学苑	http://www.nease.net/~luolian/#a2	-	-	http://bj2.netsh.com/bbs/94696/
42	中国研究	http://zgyj.freeservers.com/z.htm	+		
43	中国社会学	http://www.chinas1.com/index.asp	-	-	http://www.chinasociology.com/indexfirst1.htm
44	似乎有知识	http://zqdong.yeah.net/default.htm	-	-	http://www.cnobel.com/
45	春夏自由评论	http://www.boxun.com/freethinking/index.htm	+		
46	民主科学建中华	http://proxy.spaceproxy.com/-_-http:/dschina.com	-	-	-
47	狗眼看人	http://dogeye.easthome.net/welcome.htm	-	-	-
48	新观察	http://proxy.proxyspace.com/-_-	-	-	-

		http:/www.xgc2000.com			
49	法制经纬	http://wf-home.sd.cninfo.net/~fubai/default.htm	-	-	http://www.chinapop.gov.cn/fzjw/
50	网上文cool	http://www.cgirealm.com/articool/default.htm	-	-	-
51	胖新视点	http://go.163.com/~johnxin/default.htm	-	-	-
52	言论	http://unlimitedspeech.heha.net/default.htm	-	-	-
53	逻辑	http://luoji.126.com/default.htm	Rd: http://nujiang.myrice.com/		
54	激情如火	http://jqrh.heha.net/default.htm	-	-	http://jqrh.top263.net/
55	经济学习	http://newhome.bdinfo.net/~gyb888/default.htm	-	-	-
56	投笔从戎	http://tbcr.heha.net/default.htm	-	-	-
57	常识	http://www.bcity.com/changshi	-	-	
58	自由论坛	http://ziyou.virtualave.net/default.htm	-	-	http://www.hkfreearea.com/
59	黄焕金	http://hhj.yeah.net/default.htm	Rd: http://hxhhj.nease.net/		