Journal of East Asian Libraries

Volume 1996 | Number 108 Article 3

2-1-1996

The Year 1995: Questions about the Development of Automation in Sinological Librarianship

Thomas H. Hahn

ARTICLES

THE YEAR 1995: QUESTIONS ABOUT THE DEVELOPMENT OF AUTOMATION IN SINOLOGICAL LIBRARIANSHIP

Thomas H. Hahn University of Heidelberg

For various special reasons the year 1995, in my view, marked a turning point in the world of Chinese computing and electronic resources. Let me give a few "shining" examples to support this statement:

- The "awakening of the dragon," referring to the conceptualization and progressive implementation of high-speed fiber-optics networks[1] throughout China, linked to Japan and the United States via SprintNet, thus making it (at least in theory) possible to send files and mail to colleagues in Shanghai or Wuhan and conduct sophisticated search routines in on-line catalogs available in Beijing, Tianjin, etc., through the World-Wide Web (WWW) using beefed-up, Sinicized browsers like Mosaic or Netscape.

- TAN, the Taiwan Academic Network, goes into full operation, now providing World-Wide Web access to the Academia Sinica Institutes, the National Central Library, and a variety of other WWW servers linked together for projects and information channeling.

- Australia decides to employ a combination of INNOPAC and MASS to construct a large online repository of bibliographical and other data pertaining to library resources at the Australian National University (ANU) and other institutions with large holdings of East Asian materials. At the same time, document delivery of Chinese serials articles (108 titles) is made available by a joint effort between the Beijing National Library and ANU.

- On the software side, applications that allow full Chinese, Japanese, and Korean (CJK) language input begin to appear on the shelves: UnionWay, RisingSun, JOIN, MASS, and other software packages provide graphical interfaces to input and display Chinese, Japanese, and Korean (plus a variety of other scripts and fonts) at the same time. UNIX, DOS, and Macintosh, worlds apart only a couple of years ago, begin to approach each other in the form of LINUX running on Intel-driven personal computers (PCs), or Windows NT, implementing features that can be linked to UNIX operating systems and even extends to

[1]China now has more than twenty national networks (like CHINAPAC, CHINADDN, CERNET, etc.); on the local level mention should be made of APTLIN, the Library and Information Network of the Chinese Academy of Sciences, linking Beijing University and Qinghua University, etc.

Macintosh Powerbooks. Genuine CJK applications work on most platforms now, too, making use of the ISO 10646 norm, known to most of us as UNICODE.[2]

- The Beijing Tushuguan announced a CD-ROM product (which in fact turns out to be a digitized version of the National Bibliography) covering monographs going as far back as 1949 being available in June or July 1995 *(Zhongguo guojia shumu guangpan* 中國國家書目光盤*)*. The number of records is given as 220,000, starting with the year 1988. Update frequency is scheduled to be once a year. New CD-ROM products holding over half a million bibliographical records (including periodicals) that go back to 1949 were announced shortly after the first circular was distributed in November 1995.[3]

- This announcement followed the release of the long-awaited, important Qikan wenzhang database, first on disk, now also available on sale on CD-ROM, equivalent to the same type of bibliographical tool available from Taiwan since the summer of 1994. Rumor has it that the digitized editions of these bibliographies are much more exhaustive and complete than the printed versions. This would speak in favor of setting up CD-ROM servers everywhere and do away with the printed edition. On the other hand, update frequency is an argument against the new media, being only once or twice a year. A compromise would be to subscribe to both versions, of course.[4]

- Of interest to colleagues willing to take notice of what is going on in Europe is that the first internationally available CJK catalogs have made their appearance: in Paris at the Institut des Hautes Études Chinoises and in Heidelberg at the Sinologisches Seminar.

---

[2]For a cautious interpretation on the role and function of UNICODE, see Ken Lunde's "Online Companion to 'Understanding Japanese Information Processing'" which can be ftp-ed from ftp://ftp.ora.com/pub/examples/nutshell/ujip/doc/cjk.inf. When and how the Research Libraries Group, Inc. and the OCLC Online Computer Library Center, Inc. will move away from (R)EACC towards this more unitarian coding platform has not been announced yet, I'm afraid.

[3]Two European libraries, the Prussian State Library in Berlin and the Cambridge University Library, at present do buy Beijing MARC records to build up their bibliographical databases of Chinese book titles, but the records remain in Guojia Biaozhun 國家標準(GB) code. It would be interesting to know if there are any institutions in the US or elsewhere that also intend to buy these newly announced Beijing MARC records if only for the sake of reference, not for actually using the records to create a bibliographical database.

[4]It should be mentioned that these records which, I understand, originate in Shanghai are already following ISO 10646 (allowing for 20,902 Chinese characters), thus making quite a departure from the old GB2312-80 code (containing only 6,700 *hanzi*). I am grateful to Cathy Chiu for alerting me to the matter of completeness on the CD-ROM versus a cutback of one-third of the information with the printed edition of the *Baokan suoyin* 報刊索引.

Others have already made plans to "turn the doorknob and swing the door wide open" in 1996.[5]

- And last but not least, well-crafted homepages (some of them win prizes or make it into *Wired* or related magazines[6]) guide the curious—in vernacular script—to the most remote Japanese undergraduate school. The ration between the sheer mass of "switchboards" (i.e., homepages) we now have to access the same site of, say, the exiled Lily Wong, and the very substance these pages provide about their own environment is getting better, too.

Webpage administrators have realized that, in the long run, it is not really stimulating (nor rewarding: there is very little feedback) to live by "loaned information," i.e., just relate to a place with substantial information resources without coming up with a local representation now and again. Therefore, those managing home-spun web sites seem to have learned to approach the scholarly community to extract from them whatever the joint concept of the locality may require. In this respect I tend to see one of the most important achievements for 1995, with this long flat curve we have all been riding on for years now slowly getting ready for the liftoff. The danger that lies in wait here is that people just look at the surface of the matter, flattering themselves with nice looking graphics and fancy, formula-type searching routines for various databases, thus loosing grip on the hard substance behind the scene (or the screen). A much more knowledgeable colleague of mine lately was tempted to alert people that what in his opinion is appearing on the horizon in terms of on-line resources in approaching days does not deserve to be called WWW—this is a neutral term in any case—but, rather, should be labelled MMM, i.e., Multi-Media Mediocrity.[7]

The "highlights" mentioned above can be conveniently grouped under the headings of "information infrastructure" and "resource substance". It is trivial, but nevertheless important, to mentioned that both elements are intricately related and interwoven. In most cases still, links and substance are not pointing at the same place, nor are they provided by

---

[5]The catalog of the Institut des Hautes Études Chinoises in Paris is available (in Chinese also) using a net browser like Netscape at http://dodge.grenet.fr:8001/interroxor.html/. To access the (Chinese) Allegro-X catalog of the Sinologisches Seminar at Heidelberg telnet to sun.sino.uni-heidelberg.de, login as sinox and use the pw axopac. Item 1, access to the Twenty-five Dynastic Histories, is limited to registered scholars and institutions in Europe only. The Bodleian Library and other institutions will shortly give specifications about their online catalogs, too.

[6]*Wired*, 4.01, Jan. 96, p. 108 ff. on "The Resistance Network" aiming at Harry Wu, Burmese dissidents, etc., and again on p. 179 on Hong Kong Cinema.

[7]Courtesy of Matthew Ciolek of the Australian National University who coined the term in his article "Today's WWW—Tomorrow's MMM—the Spectre of Multimedia Mediocrity" in *IEEE COMPUTER* 29, no. 1 (Jan. 1996): 106-108.

the same agencies. With broader bandwidth (entailing faster data transmission rates) and be a matter of providing access to already existing localized resources.

Questions relating to a conference in Beijing

One form of media immediately challenging the printed media of books and serials was discussed during a conference on electronic publishing in July 1995 in China. The conference papers were put together in a volume bearing the title *Zhongguo Guji Zhengli Yanjiu Chuban Xiandaihua Guoji Huiyi Lunwenji* 中國古籍整理研究出版現代化國際會議論文集 which translates roughly as "Volume of Contributions of the International Conference on the Modernization of Publishing and Studies on the Systematizing (=*zhengli?*) of Chinese Ancient Texts" or something like that. This whole volume is one stunning parade of large scaled digitizing projects going on in Taiwan at the Academia Sinica Yuan-Ze Institute of Technology, in Hong Kong at the Chinese University, and at equally active institutions in the People's Republic of China.

For example, a project announced and discussed by the historian, Lü Zhiyi of Hebei University in Shijiazhuang, is intended to input all of the approximately 8,700 local gazetteers (*difangzhi* 地方誌) in China. With an average calculated at about 300,000 Hanzi 漢字 per title (=600,000 bytes), one single CD-ROM (storing 650 MB according to present-day standards, but maybe ten times more with the newer HDCD technology) could contain 1,100 (!) *difangzhi* in full-text format, all set for easy character-oriented retrieval, applicable to statistical counts or linguistic logics of whatever sort. The whole stock of Chinese *difangzhi* would then be available on no more than eight CD-ROMs (at present, this type of material fills whole rooms by itself), each containing about 1,100 titles. The project was described seriously in analytical terms, leaving no doubt that the project was not a balloon only filled with hot air. According to Lü, work has already begun. Note that the project description only specifies old *difangzhi* (those published before 1949) as the target material, not the now mushrooming new titles and editions.

This grand-scale project is supplemented in various ways by other institutions: the Academia Sinica, again, has not only input the twenty-five dynastic histories (about forty million characters),[8] but also many other titles, amounting to about 128 million characters and including the complete Taiwan gazetteer, *Taiwan fuzhi* 台灣府誌 , *Wenxian tongkao* 文獻通考 , the Thirteen Classics with commentary, the collection *Quan shanggu San-dai Qin-Han San-guo Liu-chao Wen* 全上古三代秦漢三國六朝文 , and dozens of other, equally important texts.[9]

---

[8]Besides the installations in Taiwan, this database can be found in the western hemisphere at Harvard, Berkeley, Seattle, and Heidelberg.

[9]A variety of institutes of the Academia Sinica are involved in digitizing projects to create full-text databases, NOT merely scanned documents!

Not quite on the same scale but nevertheless extremely useful for academics is the ICS CHANT series (*CH*ina *AN*ncient *T*exts, a project at the Chinese University of Hong Kong) which will comprise 103 titles.

The Buddhist and Taoist canons are projected to be input character by character; much of it is already beyond step one, i.e., establishing the edition the text will be taken from. Hanazono College International Research Institute for Zen Buddhism has just made available over eighty Zen texts (plus a Zen-related dictionary). A stunning project is reported taking place in South Korea where the Korean Buddhist canon seems to be well on its way to fruition; fifty million characters that will take about six to eight months to input with about thirty people working on it full-time; it is being financed by the Samsung corporation.

Many more highly interesting ventures into the binary realm of codes and bytes were discussed in Beijing. If reading the situation correctly, one could conclude that this conference volume is only the tip of the iceberg of what is really happening; many prestigious institutions were represented, but not much was said by the Chinese Academies of Social Sciences or the Chinese Academy of Science. The crucial issue for scholarly institutions and academically-oriented collections in the west is the question of accessing these types of resources. In this respect my experience leads me to say that librarians and academics will face a very uneven scene, one not much different from the Chinese book market.

1. I believe that many of the smaller databases will be available free of charge or for a very small amount over the Internet or on disk (like the *Wenxin Diaolong* 文心雕龍, the *Laozi Daodejing* 老子道德經, etc.)

2. Other more substantial and labor intensive resources will have a price tag attached to them that may run as high as $100 (texts from the ICS series) to $500 (the fancy Taiwan Hong Lou Meng 紅樓夢 electronic text edition) per text, still affordable by many institutions or private persons particularly interested in one type of text.

- There is the Taiwanese district-level *difangzhi* project;
- the New Qing History (*Xin Qing shi* 新清史, six million characters);
- parts of the Taoist collection Daozang (for example, the *Wushang Biyao* 無上必要 or Tao Hongjing's *Zhengao* 眞誥, 400,000 characters);
- the continuation of the so-called classics full-text database (Hanji Quanwen Ziliaoku 漢籍全文資料庫, thirty-five million characters);
- the daily events of modern China database (Jindai Zhongguoshi Shi Ri Zhi 近代中國史事日誌, 700,000 characters);
- the related Zhonghua Minguoshi Shi Ri Zhi 中華民國史事日誌 containing 1.6 million characters;
- and Taiwan archival material (*Taiwan Dangan* 台灣檔案, eight million characters).

3. The third category holds the most substantial electronic resources, like the Thirteen Classics with commentary, the twenty-five dynastic histories, and the Taiwan archival materials, etc. These multimillion dollar enterprises are not yet being marketed through official channels. Suffice it to say that it is extremely difficult (but also extremely rewarding) to obtain complete sets of these texts in the west, perhaps through the funding of the generous Chiang Ching-kuo Foundation.

4. The last category is the most cheerless: many of us will most probably never have access to literary or other full-text databases from the academic sector in the People's Republic of China (like the *Quan Tang wen* 全唐文 or the *Quan Tang shi* 全唐詩). They may be talked about, they may even be in operation in a given locality, but they remain closed repositories, much like many of the monastic projects aiming at inputting religious texts along the Taiwan coast.

The questions attached to this development are rather unsettling.

- For example, how are we, being in charge of more or less traditional book-centered libraries, going to cope with this massive flood of digitized data?

- How does (or can) this process influence our acquisitions policy if in, say, five years heads of acquisition departments have a choice between the digital *difangzhi* and printed ones. Which media will people go for?

- And what about the crucial public services department, the interface between the patron and the library? In a couple of years we will see people with a knowledge of digital resources equal to those beloved bibliographic maniacs that are able to relate the complete history of successive *Sibu Congkan* 四部叢刊 editions. How will this knowledge be transmitted? In library schools for East Asian resource specialists? Although I teach courses on the subject, I haven't heard of one yet.

- Again, what can we do to "educate" the researcher or the students to develop content-oriented (asking new questions about old texts?!) skills in the use of this type of resources?

- Who is going to negotiate and coordinate this type of electronic media? The person in charge would have to know computers (PCs and UNIX), the available and current editions, the software to run the whole thing. Smaller institutions will not be able to foster and keep experts of this kind.

- What sort of enterprises or joint ventures can be set up to actively initiate the construction of useful electronic dictionaries or databases, to cite only one example, related to population figures and other practical information.

You may think that all this is still a long way off. In fact, it isn't. To clothe it in a polemical robe, you could say that the "digital age" is happening right now and that

everybody has to chew off a byte or two; if not, the landscape of cogent, up-to-date scholarship will become even more fragmented due to the uneven infrastructure of networked information. It is highly desirable, and advisable, that members of academia, librarians, computer experts, and information scientists get together to discuss the options. The institutionally-backed beginning discussions that will take place in Hawaii at this year's Association for Asian Studies meeting may constitute the signal for a positive start in that new direction.[10]

[10]For instance, the Council of East Asian Libraries "Workshop on the Global Information Infrastructure" to be held on Thursday, 11 April and a related AAS panel more exclusively devoted to full-text databases called "Sinology by Computer" with participants also from the Academia Sinica taking place on Friday, 12 April.