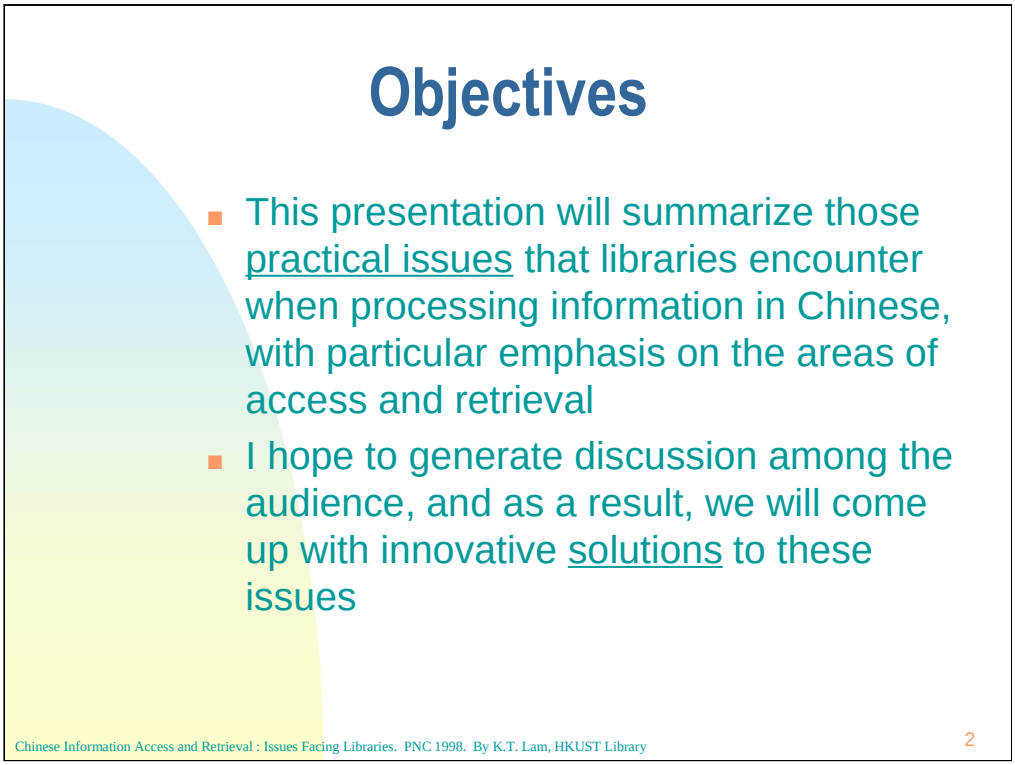
Pacific Neighborhood Consortium
1998 Conference, May 1998, Taipei

Chinese Information Access and Retrieval

Issues Facing Libraries

Ki-Tat LAM
HKUST Library

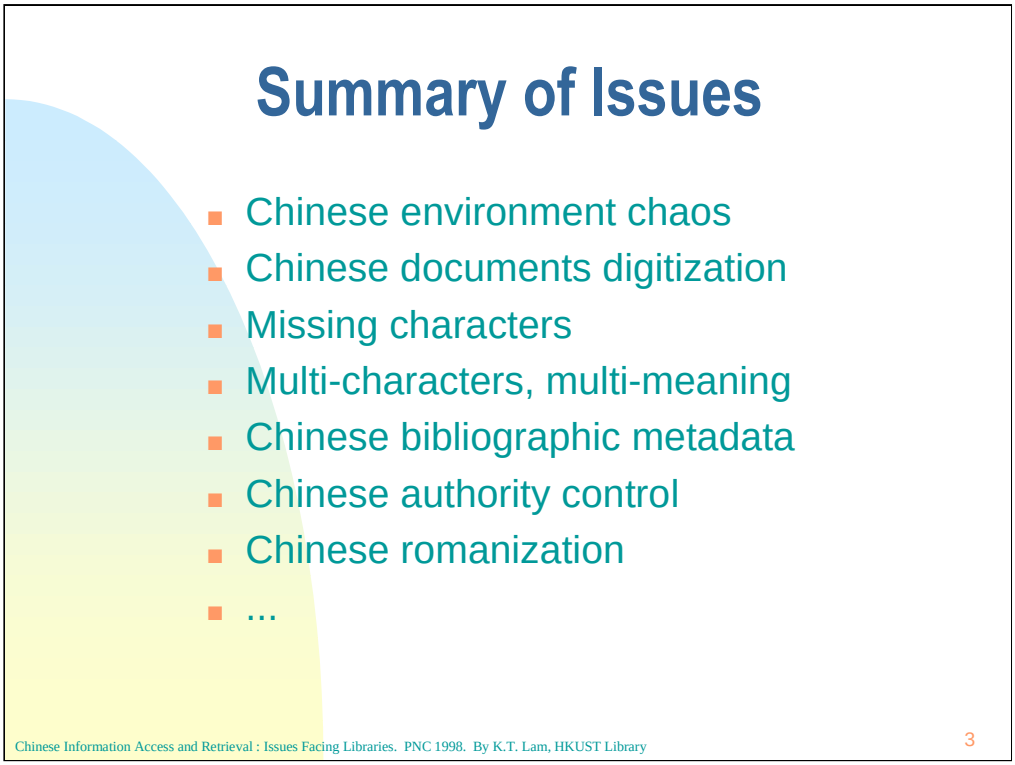


1. Chinese information processing has been very problematic in the past few decades. This is due to the fact that computer hardware, communication protocols, operating system software and application software developed so far are basically "English" oriented.

2. Libraries in East Asia have to handle information in Chinese characters and there is an immediate need to resolve the outstanding issues. This presentation will summarize those practical issues that libraries encounter when processing information in Chinese, with particular emphasis on the areas of access and retrieval.

3. Although this presentation focuses on Chinese scripts, it is likely that multilingual libraries using non-roman scripts (including Japanese, Korean, Thai, Russian, etc.) are also facing the same issues.

4. It is hoped that this presentation will provide the background information needed for discussion and problem solving.

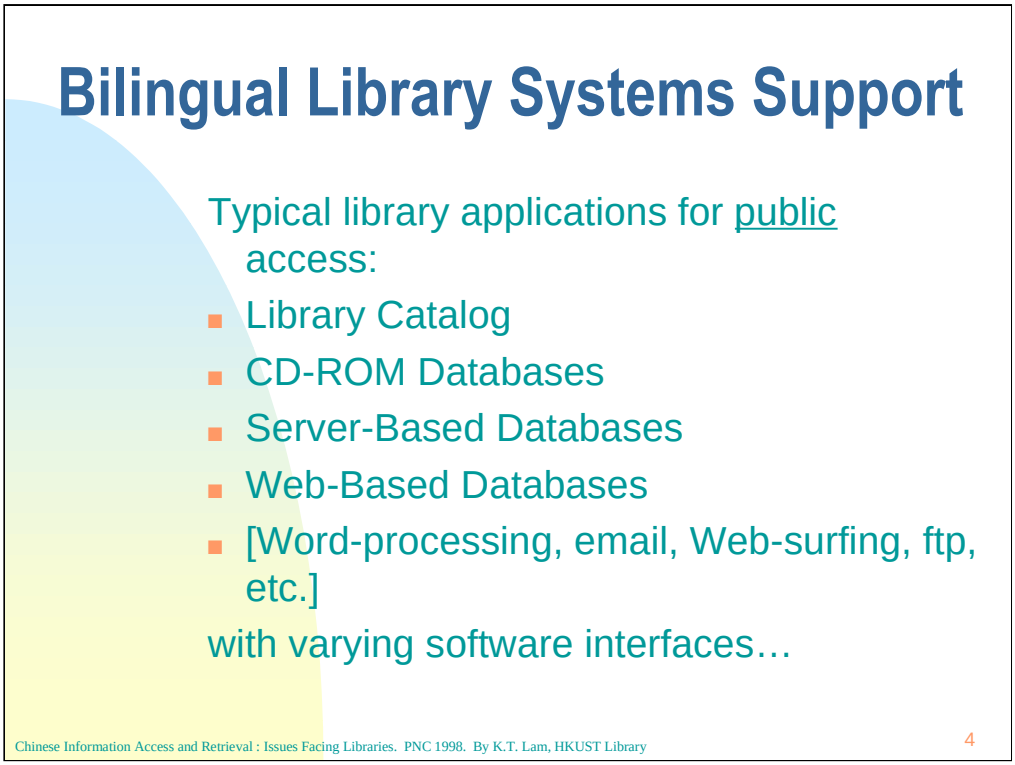


1. We will start the presentation with general issues in Chinese encoding and software localization and explain how these issues have caused so many problems in supporting library systems that are bilingual in nature.

2. Then we will look at two Chinese language specific problems, i.e. the handling of missing characters and the issues of muti-characters and multi-meaning.

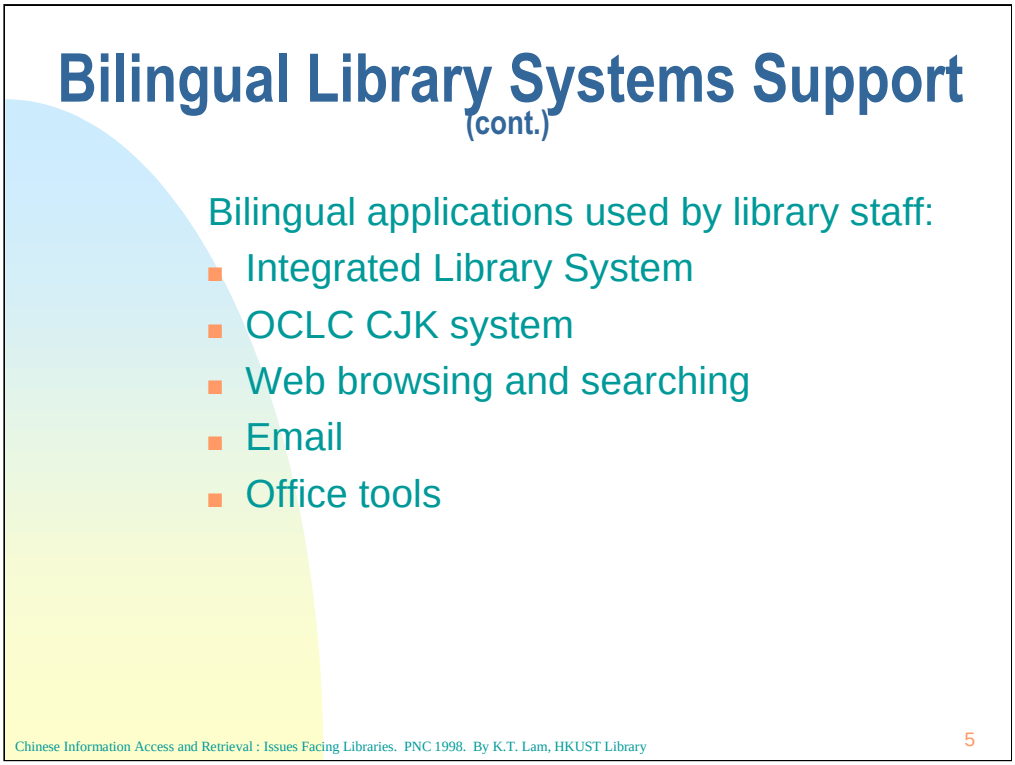
3. We will then discuss issues that are specific to the library catalog, including bibliographic metadata format, authority control and Chinese romanization.

4. Due to time constraints, other related issues will not be discussed in this presentation. These include problems in machine translation, Chinese collation (sorting), Chinese phase indexing and fast searching algorithm, etc.

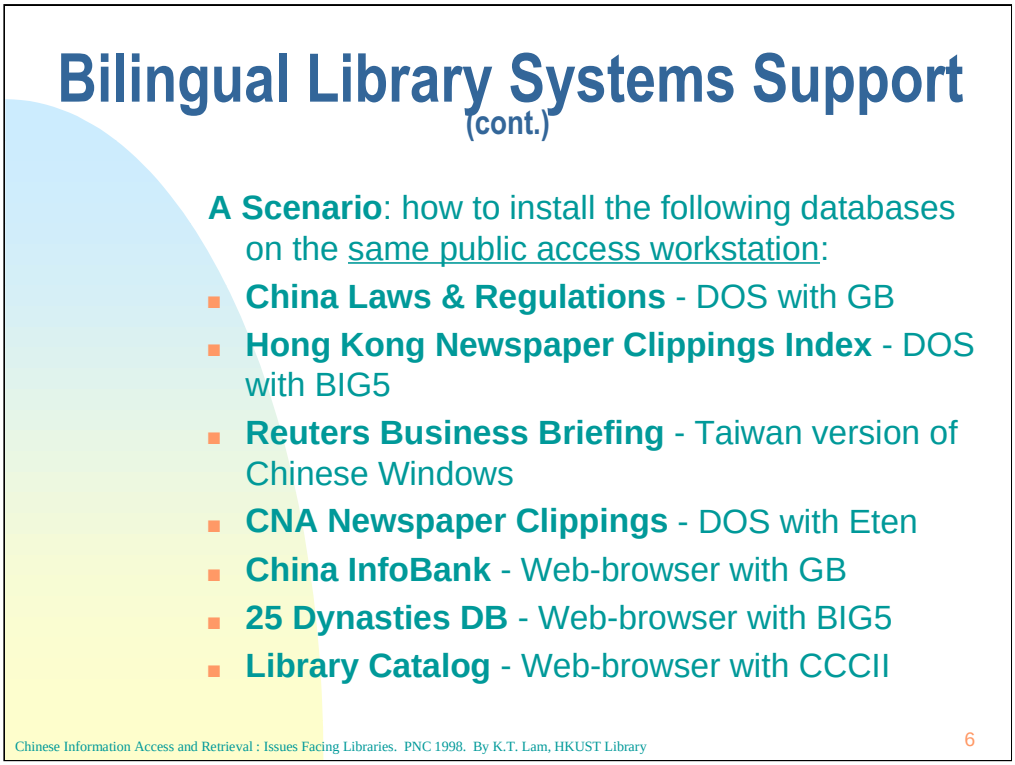


1. HKUST has a strong collection of electronic resources. They are mounted on the Intranet for network access. Some of these databases are in Chinese.

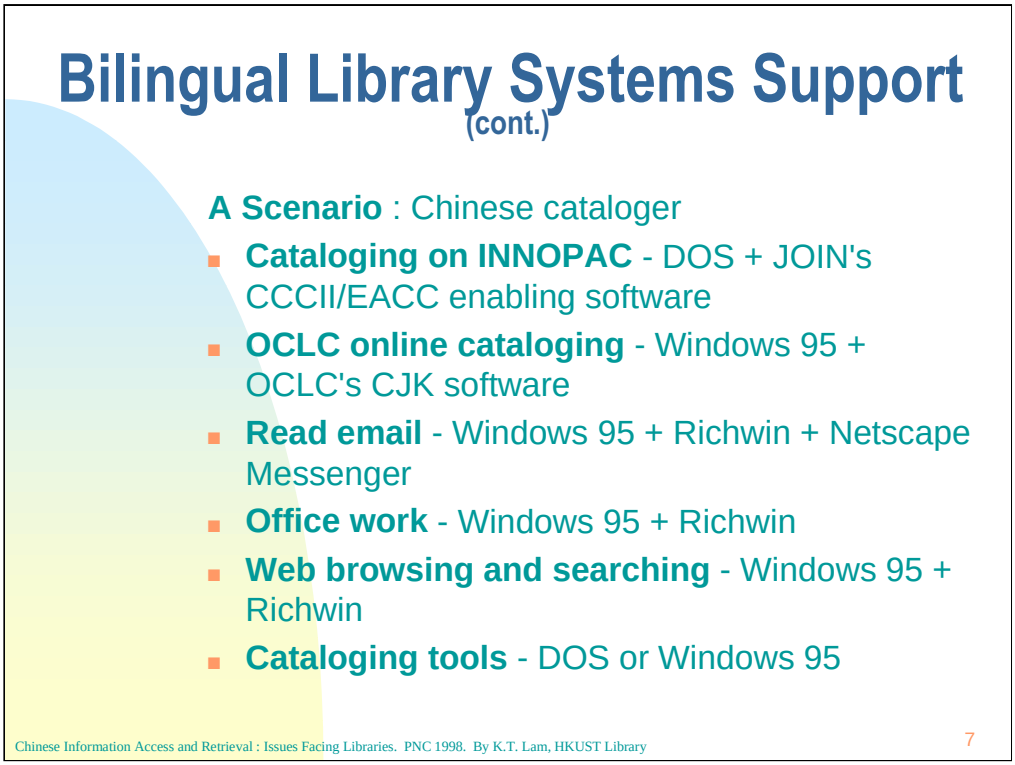
2. It is not trivial to install, integrate and maintain this Library Online Systems environment because it involves different kinds of server platforms, protocols, and search interfaces. The support of a bilingual computing environment for public access is particularly problematic.



1. In addition to providing public access, the Library Systems Department also needs to support the bilingual computing environment for library staff access.



1. It is extremely difficult, if not impossible, to install all the above databases on the same machine.

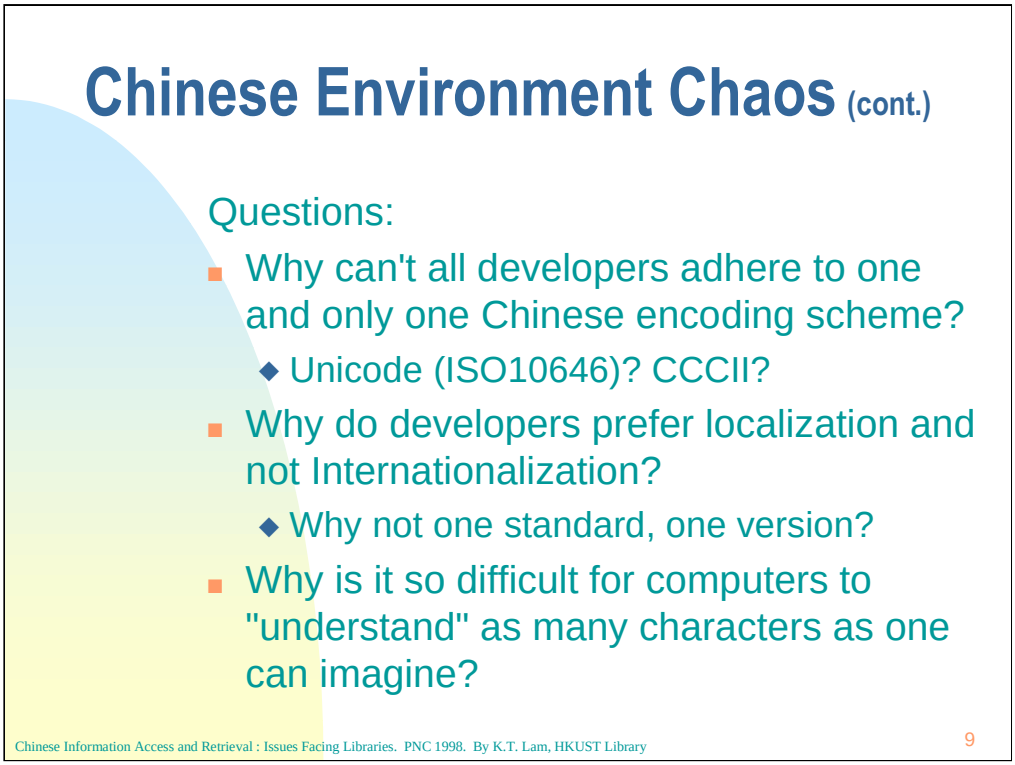


1. Note that this cataloger has to reboot his/her machine between DOS and Windows 95 quite a number of times each day!

1. It is unfortunate that there are so many Chinese character encoding schemes and that different regions prefer different schemes.

2. It is also unfortunate that OS and application vendors prefer to "localize" their products to a particular encoding scheme. As a result, libraries have to support multiple and incompatible Chinese computing environments.

3. Note that UTF8 and Unicode are still not common for library uses.

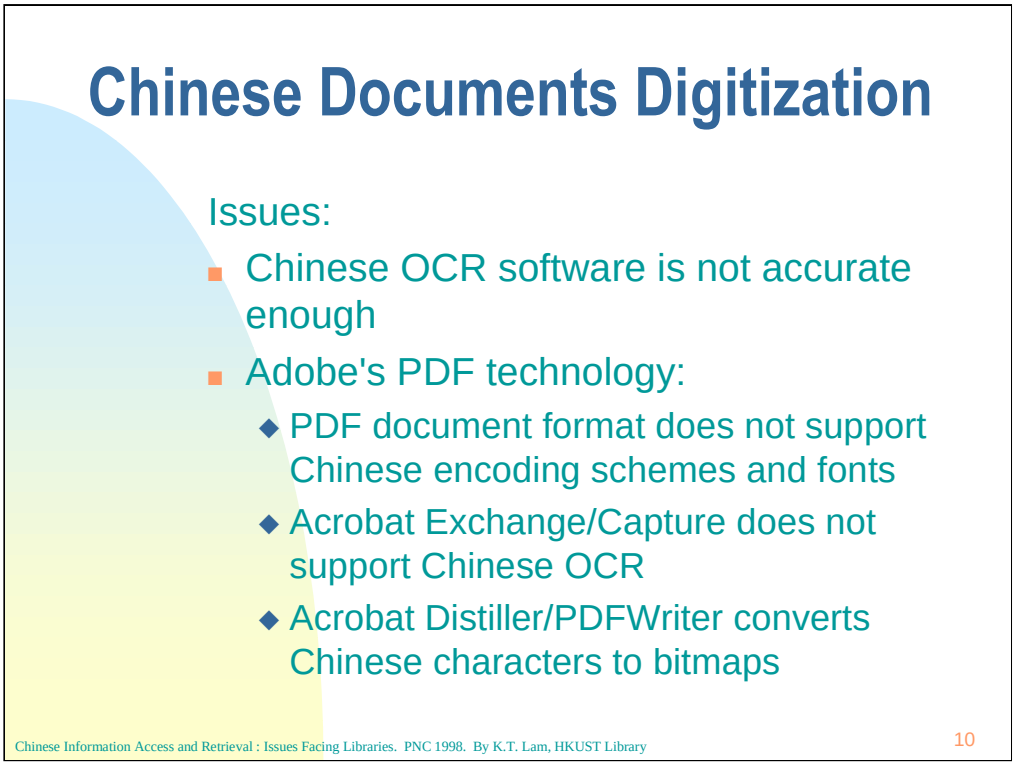


1. Life will become simpler if all Chinese library applications run on one and only one Chinese environment. The ideal Chinese environment will have a large character set, and will be popularly supported by software vendors and users.

2. Internationalized software supports any language without cultural contents. Localized software is a customized product that supports a particular language (locale), and is therefore culturally focused.

3. Microsoft's strategy has been to localize its OSs by creating different versions for different regions. The result is that data created on one version cannot be seen on another version. This is still true for Windows 95 and NT, although they try to reduce the effect by making their applications unicode-based.

4. CCCII is certainly a better character set for libraries. However, we see that Unicode is getting more and more popular. Is it possible for library systems to use CCCII as the internal character set for storage and use Unicode (or UTF8) as an external encoding scheme for processing and interchange?

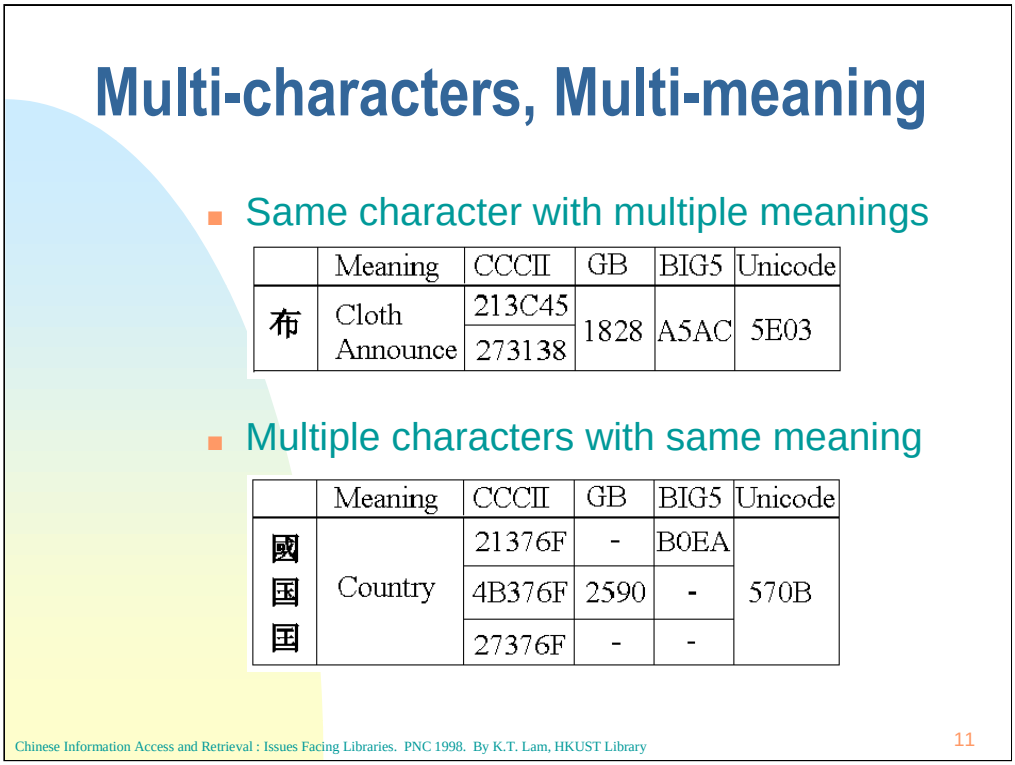


1. HKUST Library has been digitizating documents, including University Archives, theses, course related materials, and lecture notes. So far, digitization of Chinese documents is the biggest headache to us.

2. It is difficult to find Chinese OCR software that can accurately recognize various Chinese fonts and at various sizes. It is essential that the software is also able to recognize English.

3. HKUST Library adopted Adobe's PDF file format for the various digitization projects. However, PDF format does not support Chinese encoding schemes and fonts. In addition, Acrobat tools are not "internationalized" or "localized" to handle non-roman documents, such as Chinese documents. The "work-around" is to store Chinese in the form of bitmaps in PDF.

4. Adobe's failure to "internationalize" its Acrobat products and PDF format shows the fact that many software and standards developers are still very short-sighted on the importance of handling multilingual scripts in the modern software market.



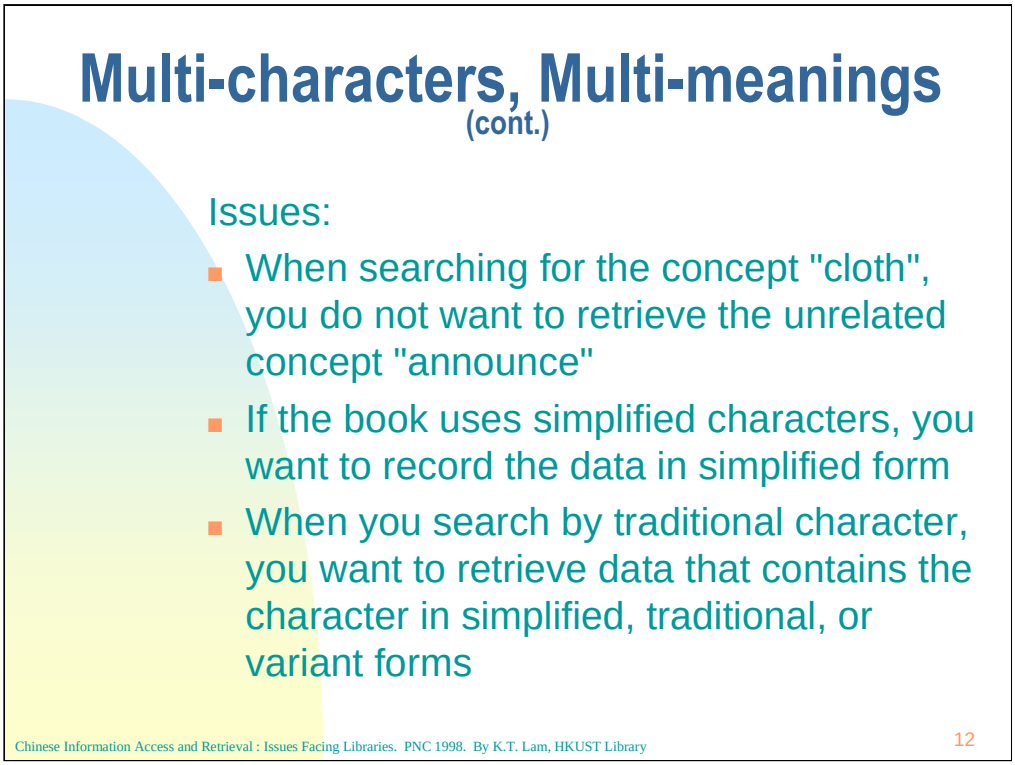
1. Chinese character encoding schemes deal with the "shape" (glyph) of the character; no semantic attribute is involved.

2. Due to historical reasons, many Chinese characters' shapes were simplified and/or modified. However, libraries have to preserve the original forms recorded in books.

3. The examples above show two different aspects of the problem of retrieval due to the one-to-many and many-to-one mappings between character (glyph) and meaning. (To be explained in next slide.)

4. Note that CCCII allocates different coding points for different meanings and different "shapes" (glyph). This information is very useful to retrieval software in differentiating the desired meaning and shape.

5. Also note the "unihan" concept in Unicode (and ISO 10646), i.e. one and only one coding point is allocated to all "glyphs" of the same "character".



1. In the case of multiple meanings:

- how to preserve the meaning of the character in the database, and

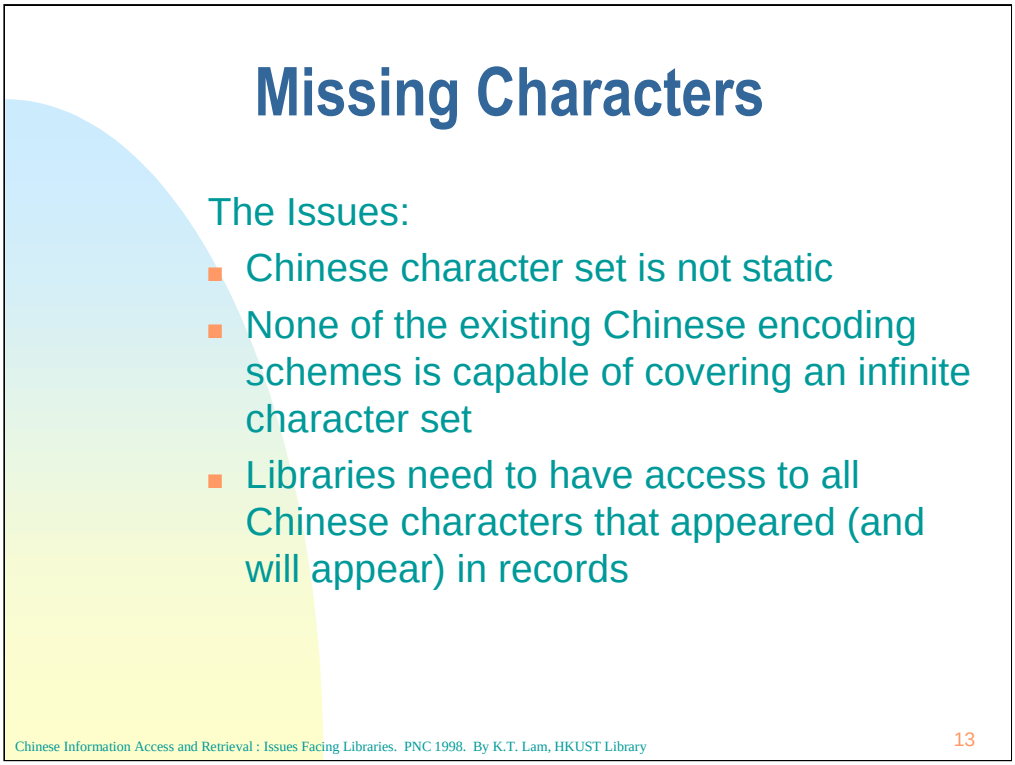
- how to make the retrieval system know which meaning you are referring to when you input such a character in the query.

2. In the case of multiple character shapes (glyphs):

- how to preserve the glyph information in the database, and

- how to retrieve all forms that are related to the glyph you input in the query

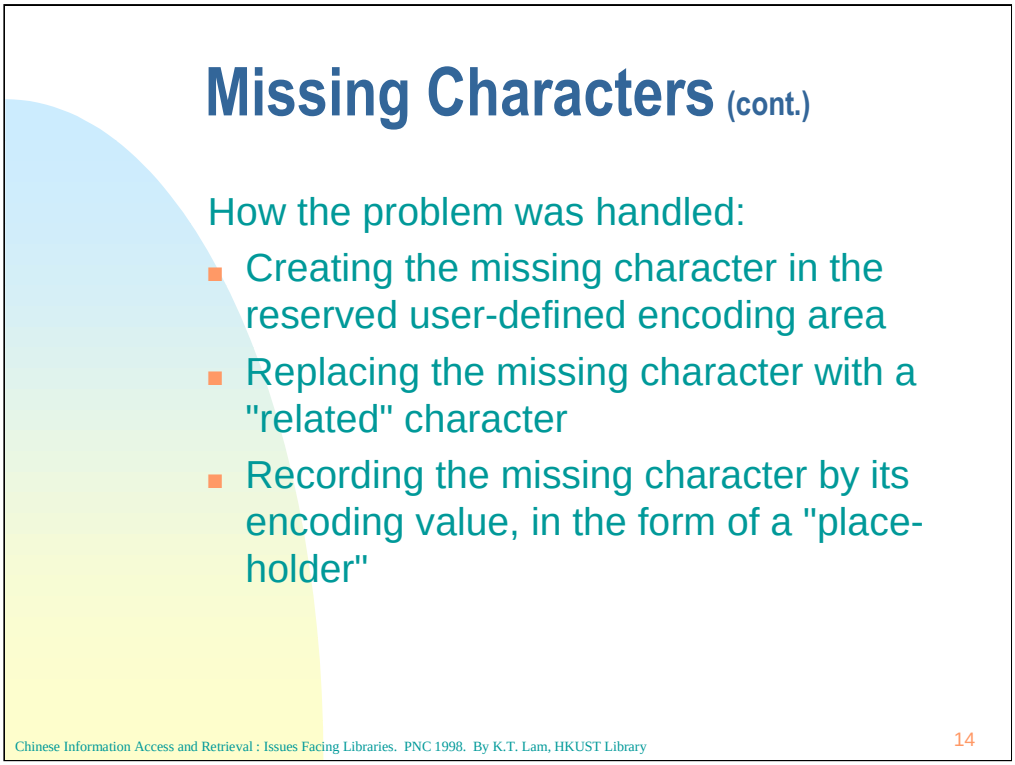
3. Solution welcome!



1. The Chinese character set is an infinite set; its growth and changes are a function of time. It is difficult, if not impossible, to freeze its growth.

2. Character encoding schemes, e.g. ASCII, BIG5, Unicode, etc. are basically standards to facilitate the interchange of popularly used ("modern") characters. They do not bear the responsibility of preserving historical records.

3. According to the cataloging practices, catalogers are required to transcribe the exact "shape" of the characters, e.g. they cannot change a simplified character to its traditional form; or change a Kanji form to a Chinese form.



1. HKUST Library implemented all the above methods in the course of the past 7 years.

2. At first, we created more than 2000 simplified characters in the BIG5 user-reserved area, and implemented them on Eten. We needed to do so at that time because there was no Chinese enabling "software" that supported both simplified and traditional characters simultaneously.

3. INNOPAC (the integrated library system at HKUST) uses EACC/CCCII as the internal character code for storage. When displaying on a BIG5 encoding screen, missing characters will be mapped to their traditional form (thanks to EACC/CCCII's linkage information) and converted to BIG5 on-the-fly (based on a conversion table). If the conversion fails (not found in the conversion table), the missing character will be displayed as an ASCII string of its EACC/CCCII encoding value.

4. It should be noted that the above methods do not resolve the missing characters thoroughly.

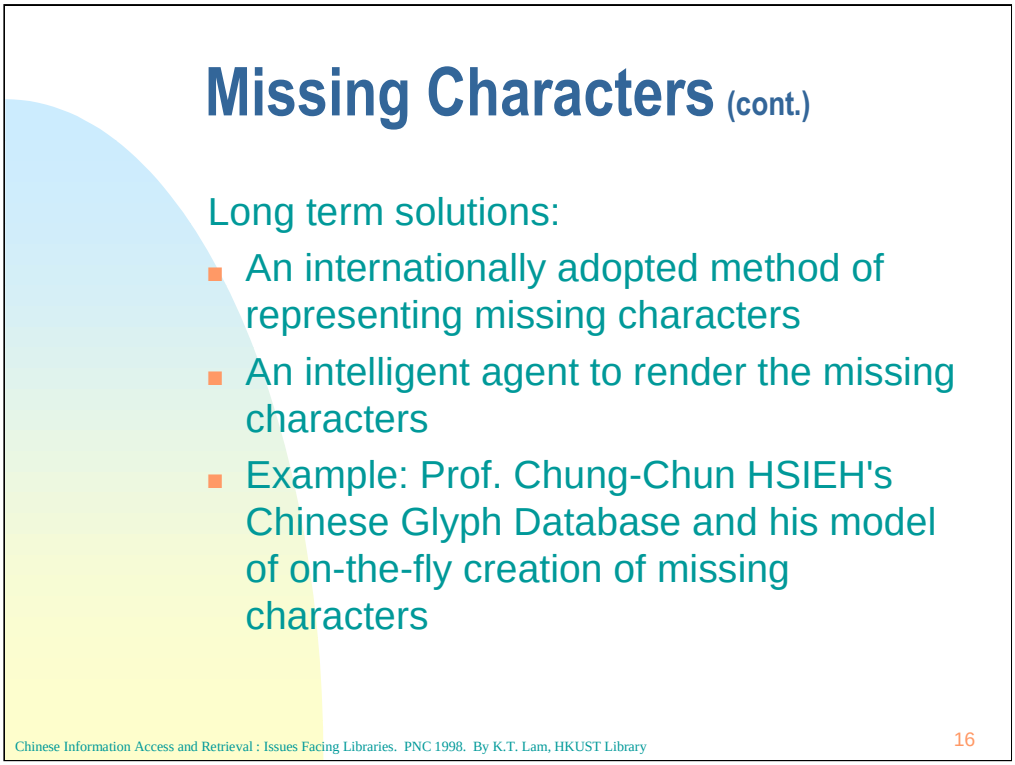
How INNOPAC handles missing characters

鉴

A missing character represented by the ASCII string of its CCCII encoding value

国

A simplified character was replaced by its traditional form



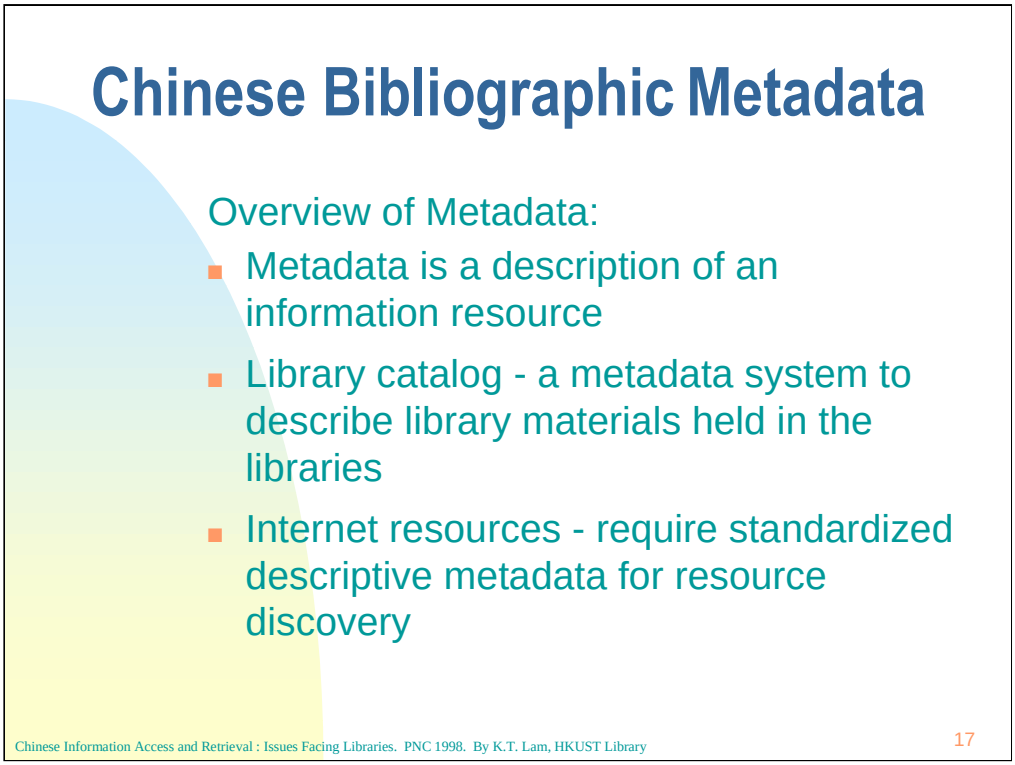
1. An international standard is urgently needed to represent missing characters for information interchange. These include how to embed missing characters in HTML/SGML/XML, in storage and retrieval systems (e.g. library catalogs), and in popular document formats (e.g. Word, PDF, Postscript).

2. Information embedded will be used by an intelligent agent for rendering.

3. The role of the intelligent agent is to provide facilities for rendering the missing characters at the client end for displaying, inputting, outputting, searching, and storage.

4. The intelligent agent may work with a character database containing rules and attributes for the rendering.

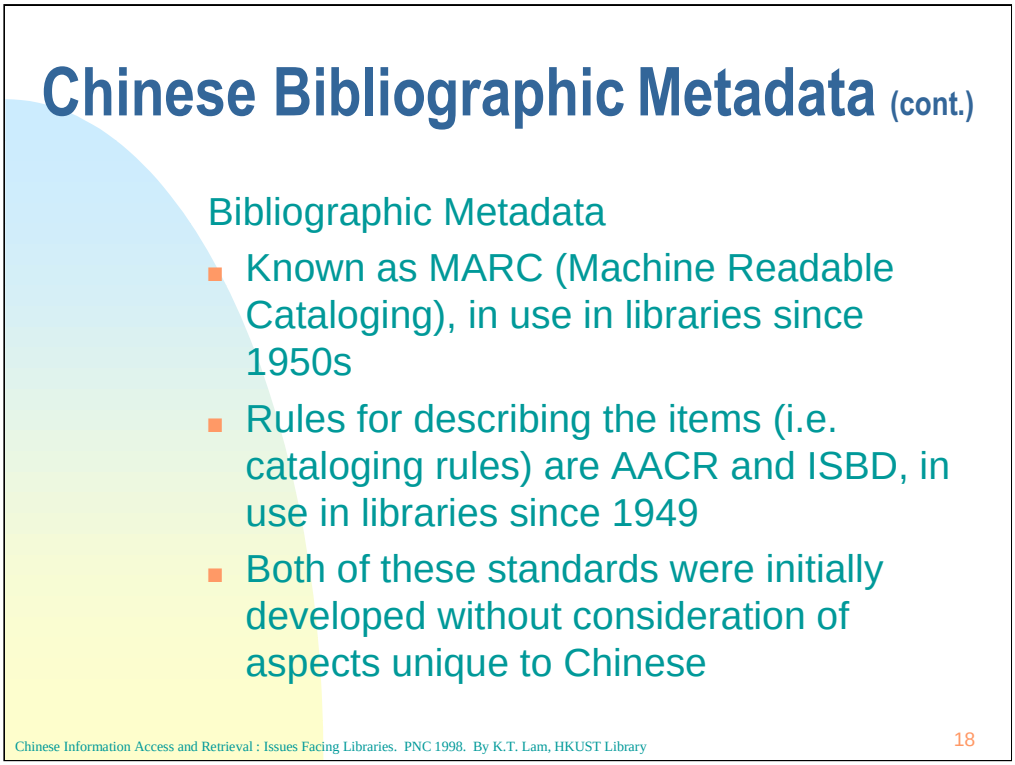
5. It is important that the intelligent agent (in the form of a set of API) and the character database be made publicly available, so that software and system developers can incorporate the technology into their products.



1. Metadata is data that describes the origins and the use of data.

2. Bibliographic (cataloging) data is a kind of metadata that describes library materials. Elements of bibliographic data include author, title, imprint, subject, call number, etc.

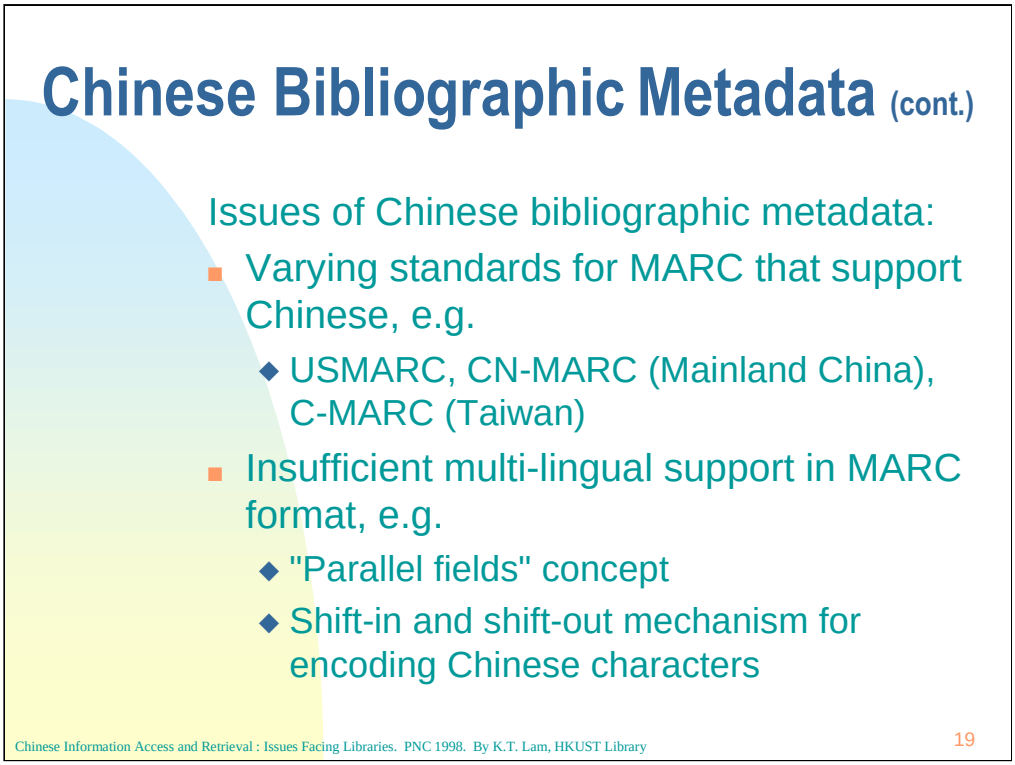
3. "Although the concept of metadata predates the Internet and the Web, worldwide interest in metadata standards and practices has exploded with the increase in electronic publishing and digital libraries, and the concomitant "information overload" resulting from vast quantities of undifferentiated digital data available online." [extracted from http://128.253.70.110/DC5/UserGuide4.html]



1. MARC format has been widely used in the library communities for bibliographic data interchange and for cataloging. Most of the commercially available library automation systems support the MARC format.

2. AACR2 Revised and ISBD are also widely used in the library communities.

3. As these standards were initiated in the Anglo-American communities, opinions unique to Chinese and other non-roman scripts were not well represented during their development.



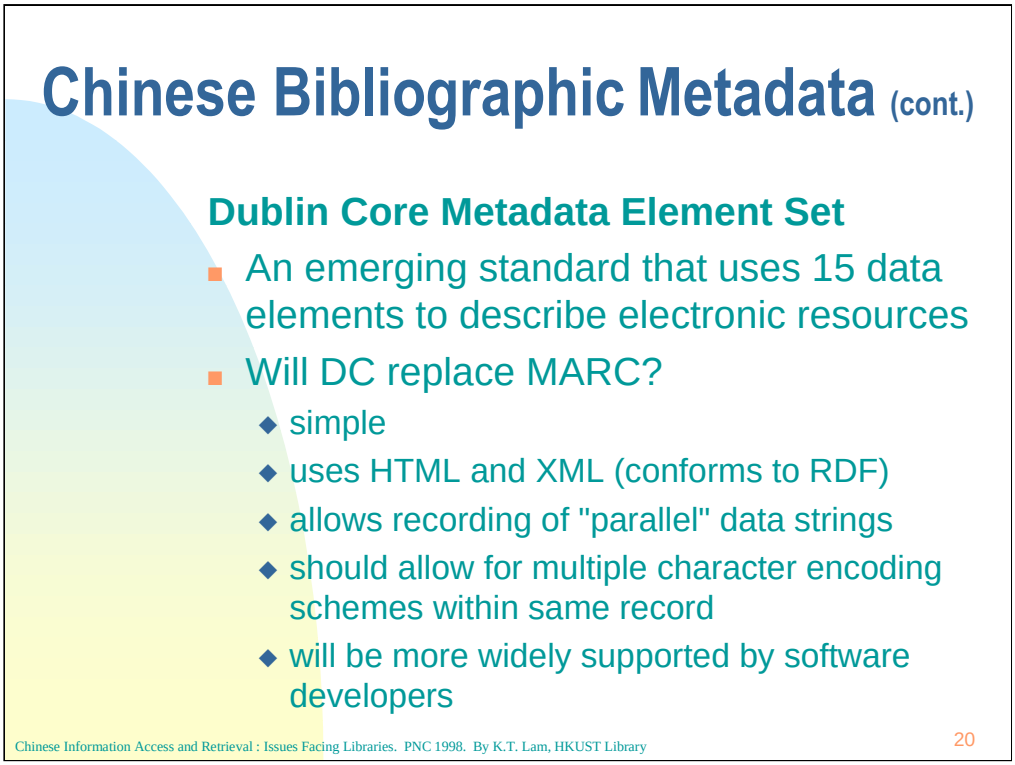
1. ISO 2709 specifies the MARC standard for interchange of bibliographic data. However, different regions have developed different standards for content designation. This is very similar to a situation where mutliple DTDs of the SGML standard were developed to define the same document type.

2. It is difficult for bibliographic data to be interchanged between Hong Kong, Mainland China, and Taiwan, because their MARC formats are different from each other and therefore incompatible.

3. In the early USMARC standard, only roman data was allowed. In order to accomodate CJK data, a tag called 880 was added to the standard, acting as a "parallel" field to its romanized form. This parallel mechanism is very difficult to implement on library automation systems.

4. It should also be noted that the parallel mechanism in USMARC only allows for linkage between two data strings. However, there are situations that require linkage for more than two data strings, e.g. data strings in CJK characters, Pinyin, Wade-Giles, and Japanese romanization.

5. USMARC uses shift-in and shift-out escape sequences to encode Chinese characters -- very cumbersome.



1. There are a number of emerging metadata standards for electronic resources, including TEI (Text Encoding Initiatives), AHDS (Arts and Humanities Data Service), SOIF (Summary Object Interchange Format), and Dublin Core.

2. Dublin Core specifies 15 elements (Title, Subject, Description, Source, Language, Relation, Coverage, Creator/Author, Publisher, Contributor, Rights, Date, Type, Format, Identifier) for describing electronic resources.

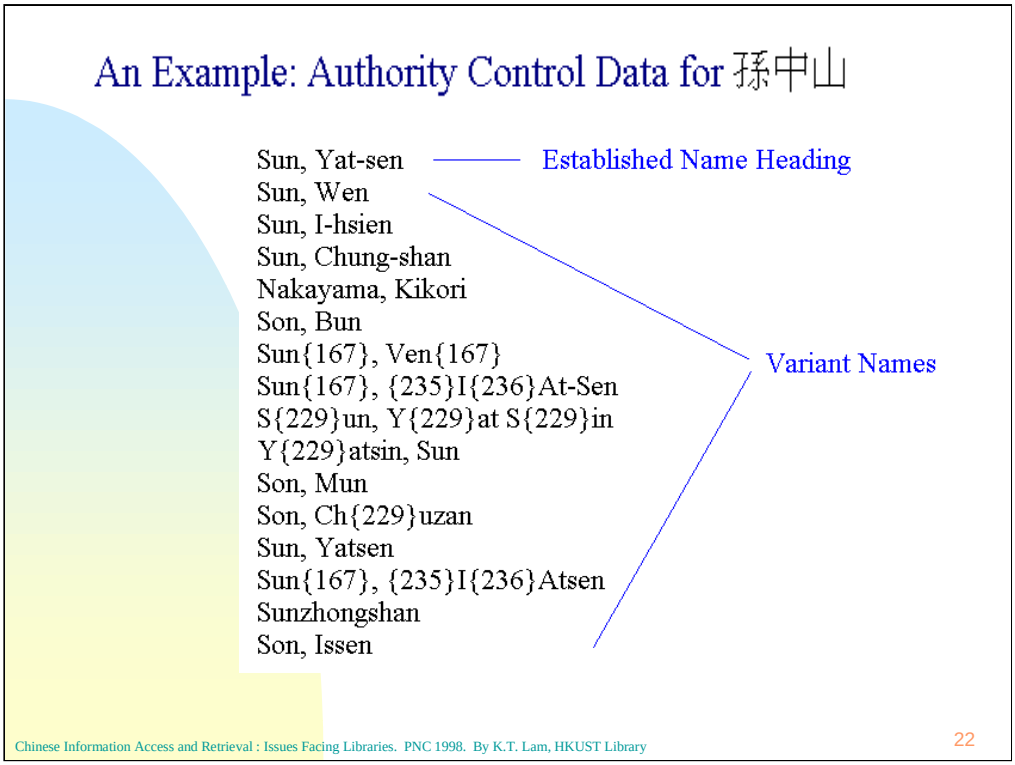
3. When catalogers are talking about minimal or semi-full MARC cataloging to reduce the backlog, Dublin Core seems to be a logical choice, as it is much simpler than MARC format.

4. However, are the people involved in developing the Dublin Core aware of the issues involved in Chinese scripts? We obviously do not want this standard to follow in the same footsteps as the development of MARC.

Chinese Authority Control

Overview of Authority Control:

- Authority control is an effort to standardize the use of headings (i.e. names, uniform titles, and subject headings) in the bibliographic metadata for better retrieval
- Authority control data (that describes the use of a heading) is established by authoritative agents, such as the Library of Congress



1. This is the authority control data for "Sun, Yat-sen" established by the Library of Congress (LC). Note that a large portion of libraries around the world use LC's authority control data.

2. However, LC does not support Chinese scripts in its authority control data. It is illogical to use a romanized form of a Chinese name as the established heading.

3. In the above example, there is no reference to the name in Chinese characters. It is difficult to determine which person this authority control data refers to.

4. These have been causing a lot of trouble for Chinese catalogers. On the one hand they love to follow LC's standard, but on the other hand they are forced to modify the authority control data supplied by LC by adding the Chinese characters to the record. A Chinese author's real name (in Chinese characters) is surely more authoritative than a romanization!

1. To make the situation worse, the authority control data format (also in MARC format) is incapable of handling "multiple established headings", although this is so common in Chinese names.

2. Since there is no standardized way of handling Chinese authority control, no library automation system supports the retrieval of Chinese names satisfactorily.

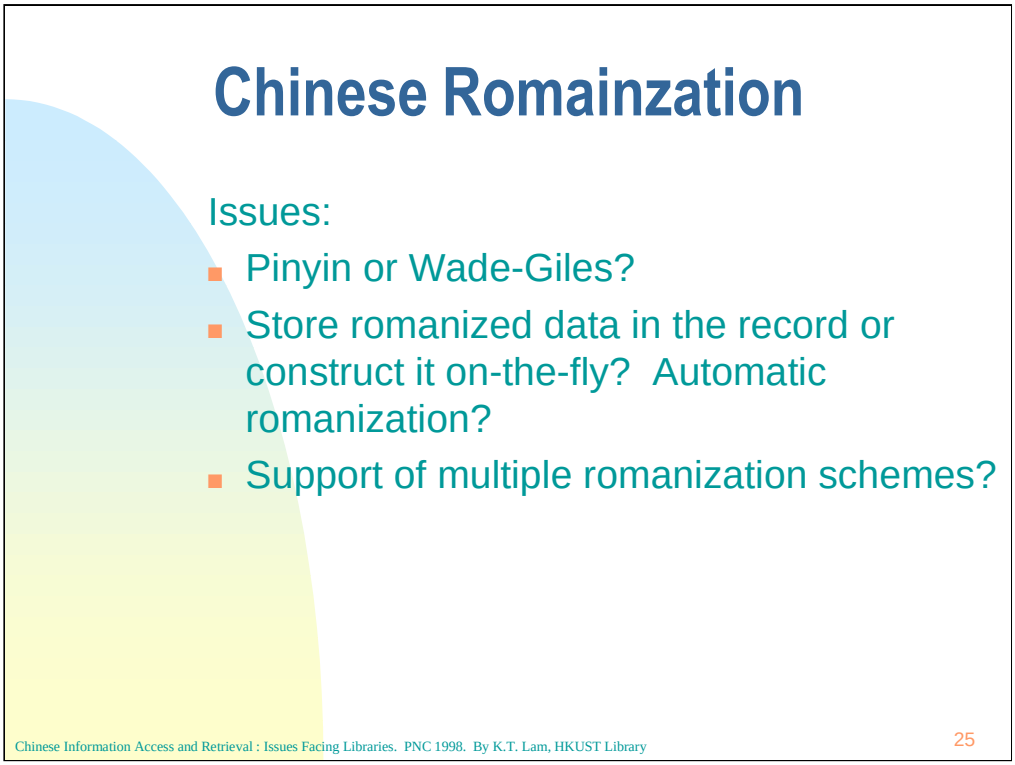
Chinese Authority Control (cont.)

The issues:

- No standardized Chinese authority control data format
- No authoritative agency that issues Chinese authority control data

I would like to see the establishment of a task force that comprises experts from the Mainland, Taiwan, Hong Kong, and overseas to study problems related to Chinese authority control.

Can the Pacific Neighborhood Consortium take the lead?



1. The Library of Congress has just announced that it will adopt the Pinyin system of romanization of Chinese by the year 2000. Planning is well under way.

2. Problems anticipated in the switch from Wade-Giles to Pinyin are:

- substantial effort needed to convert retrospective records from Wade-Giles to Pinyin

- the argument over whether libraries should adopt Chinese romanization rules and should connect or aggregate individual Chinese syllables.

3. If romanization data can be automatically generated, then libraries do not need to store it in the bibliographic and authority records. Before automatic romanization is possible, we need to resolve the one-to-many problem and the aggregation problem.

4. It should be noted that a Chinese data string can be romanized in many different romanization schemes, including Pinyin, Wade-Giles, or the Japanese or Korean romanization schemes. We need a computing environment that supports multiple romanization schemes.

Conclusion

Requirements:

- Study on the coexistence of CCCII and Unicode for library use
- Developers should produce internationalised software
- Establishing a mechanism for rendering missing characters
- Research into the retrieval issues related to data in Chinese scripts

Conclusion (cont.)

- Metadata standard developers should be aware of issues specific to Chinese scripts
- Establishment of a task force to tackle the issues in Chinese authority control
- Research into the concept of automatic romanization

Thank You!