

Building a Global Network of Chinese Genealogical Information

建造一個全球性的中華譜牒信息網絡

Bin Zhang (Vice President of DAZUPU; California State University, Sacramento)

張斌 (中華大族譜協會副理事長 , 加州州立大學薩科拉門多分校圖書館)

The mission of DAZHONGHUA Family Tree Development Association (a.k.a. DAZUPU) is to develop a global repository of genealogical information of Chinese descendants. This paper addresses some technical issues related to this family tree project.

中華大族譜協會 (英文簡稱 DAZUPU) 的宗旨是開發-建造一個存儲全球中華譜牒信息的數據庫。本論文將介紹與此項目有關的一些技術問題。

Scope of the Project: 該項目的內容

It is not our goal to digitize (i.e., scan) the original genealogical records and provide access in its original form. The end product of the DAZUPU project is a web-based user interface that allows access to genealogical data of all Chinese families worldwide. A user will be able to search, browse the database, and generate an electronic family tree based the user's query and access rights. Authorized users will be to add new data or update existing information via a web interface, or load data in batches.

把譜牒資料原件簡單地數字化 (掃描) 讓用戶能進入原有的格式 , 並不是我們項目的目的。我們的最終目標是為用戶建立一個以互聯網為基礎的界面 , 讓他們能從世界任何地方進入所有中國人的族譜。獲得授權的用戶可以從中搜索 , 瀏覽該數據庫並按用戶要求和權限建立電子家譜樹 , 還可以添加新信息、更新既有的信息 , 或成批導入資料。

The project involves these tasks:

- Identify, adopt or develop international standard(s) for metadata collection and exchange
- Collect and process collected genealogical data
- Catalog genealogical records at individual and family record levels using established metadata schemas
- Implement genealogical information management system(s) that supports the above standard(s) and provides access to the data at all levels.

該項目包括如下任務 :

- 確認 , 採用或開發一個收集和交換元數據 (見下文) 的國際標準

- 收集和處理譜牒資料
- 在個人和家庭記錄的層次上對譜牒資料進行分類
- 建立譜牒資料的管理系統，使之能支持上述標準並可進入所有的層次

Metadata 元數據

Metadata is a fancy term to mean the data that is used to describe other data or information sources. Metadata that describe a book, for example, could include such information as title, author, editor, publisher, summary, call number, etc. These metadata elements can be used to describe a person in a genealogical information system: name, gender, date and place of birth, nationality, phone number, residential address, e-mail address, etc.

元數據是一個時尚的術語。用它來描述數據資料和信息資源。例如，我們可以用包括書名，作者，出版者，內容摘要和書號等數據元來描述一本書，用姓名，性別，出生日期和地點，國籍，電話，住址，電郵郵址等元數據元來描述個人，等。

To build a global Chinese genealogical information system, it is critical that all the metadata is collected/generated using the same standard.

建立一個世界性的中國族譜信息系統的關鍵，是對所有元數據的採集和產生使用同一個標準。

Libraries and/or archives have been collecting family genealogical records for years. These records have been cataloged and indexed the same way as books using well-established bibliographic metadata standards used in the library field, such as **MAchine Readable Cataloging (MARC)**, **Dublin Core (DC)**, etc. Metadata standards for describing individual persons and/or their immediate families (at individual level), which are critical in genealogical information systems, however, are not as well established. **GEnealogical Data COMmunications (GEDCOM)**, though not an official standard, is the *de facto* standard that is well-supported by genealogists and genealogical software vendors.

圖書館和/或檔案館收集家譜信息已多年，這些資料就像其它印刷品一樣被作為數據源分類，用的是圖書館業界早期建立的元數據標準，如 **MAchine Readable Cataloging (MARC)** 和 **Dublin Core (DC)** 等等。

可是，對於人物個體以及他們直系家人的元數據標準並未好好建立。儘管 **GEnealogical Data COMmunications (GEDCOM)** 並非正式的標準，但它已被譜牒工作和譜牒軟件製作者普遍接受。

GEDCOM (譜牒元數據通用標準)

GEDCOM was developed and has been maintained by the Family History Department of Church of Jesus Christ of Latter Day Saints (LDS). The last official GEDCOM edition is 5.5. GEDCOM 6.0, the first XML-based version was proposed in 2001, but has not become official. In the mean time several other XML-based proposals have been drafted by other individuals and/or organizations. So far none has been recognized by the industry as the standard that would replace GEDCOM 5.5.

GEDCOM 是由耶穌基督聖徒教會 (LDS) 的家史部開發的。最新的版本是 5.5。GEDCOM 6.0 是第一個以 XML 為基礎的版本，2001 年推出，但尚未正式發布。與此同時，有其他個人和組織也推出多個以 XML 為基礎的版本。但至今仍未有一個被業界認可取代 GEDCOM 5.5。

Here are some benefits of GEDCOM: 這裡列出一些 GEDCOM 的優點：

- Support by vendors and the industry 軟件供應商和業界的支

Many online genealogical databases allows user to search and download data in GEDCOM format. Examples of such services include FamilySearch.org and RootWeb.com. One can easily download all records that he/she found in either system as GEDCOM file, and import into a genealogical information management system (GIMS) of his/her choice (as mentioned earlier, GEDCOM is supported by almost all GIMS).

許多在線的譜牒數據庫(服務系統)允許用戶以GEDCOM的數據格式去搜索和下載。例如FamilySearch.org 和 RootWeb.com. 從它們中任一個系統你都非常容易下載你找到的所有的以GEDCOM格式儲存的 (譜牒) 記錄 (資料)，並把它輸入你所選用的任意個譜牒信息管理系統 (GIMS) (請注意：我們早前已經提及GEDCOM 基本上是被所有的GIMS支持的。)

- UNICODE Support UNICODE 支持

UNICODE is one of the encodings supported by GEDCOM. This is an important feature for Chinese genealogical data.

UNICODE 是 GEDCOM 支持的解碼/譯碼系統之一。這對中華譜牒資料是一個重要的性質。

- Support of Multimedia Objects 多媒體物象的支持

GEDCOM Supports attachment of multimedia objects, either linked or embedded, to individual or family records. This feature allows us to enhance the traditional family records by adding photos, audio or video clips, links to websites, etc. to a individual or family records or to a particular events.

GEDCOM 支持多媒體附件對人物個體或家庭記錄的嵌入或鏈接。這個性質讓我們可以增強傳統的家譜記錄，把相片，音像，錄影等連接到網站上，添加到個人或家庭的記錄或事件中。

- Support for User-defined Elements 支持用戶自定義元素

GEDCOM allows user-defined data elements. This allows us to include information that is unique to Chinese family genealogical data.

GEDCOM 還容許用戶自定義數據元素。這就讓我們可以用它去記錄中華譜牒資料獨有的信息。

有關的GEDCOM和其他元數據標準的更多信息，請訪問DAZUPU的網站：<http://www.dazupu.org>

At this time, we have identified GEDCOM 5.5 as our standard data exchange format for DAZUPU project. It is important to note that we are not recommending that GEDCOM be used as the internal data structure. Rather, all the systems that will participate DAZUPU project need to be able to read and export records in GEDCOM format without data loss. We will continue to watch the development of other XML-based metadata schemas.

目前，我們確定 GEDCOM 5.5 作為我們 DAZUPU 項目的標準數據交換格式。要注意的是，我們不建議把 GEDCOM 用於內部數據結構。儘管所有參與 DAZUPU 項目的系統需要用 GEDCOM 格式去閱讀和輸出記錄才不會丟失數據。我們將繼續關注其他基於 XML 的方案開發。

GEDCOM was originally developed for the Church of Jesus Christ of Latter Day Saints (LDS). There are many data elements in the specification that do not apply to Chinese families. On the other hand, there are data elements that are common in Chinese genealogy but not covered by the GEDCOM specification. GEDCOM permits user-defined data elements. We need to identify these elements and add to DAZUPU GEDCOM guidelines.

最初是為耶穌基督聖徒教會 (LDS) 建立的 GEDCOM 規範，有許多數據元素並不適用於中國家庭 (中華譜牒)。另一方面，有些中國家庭記錄普遍需要的數據元素，並不在 GEDCOM 規範之內。我們需要確定這些元素，並正在添加到 DAZUPU 的 GEDCOM 使用指南裡。

***Data Collection* 數據採集**

Libraries and archives have been collecting and cataloging family genealogical records for years. But the scope of the data collected by them has been limited to information about the records themselves (source level). Little information about individuals in the family, except brief description about a few notable individuals is recorded in library catalogs. The focus of the project is to extract the information contained in Chinese family records (家譜) and store in a well-structured electronic format. The information here include basic information about individuals (such as names, gender, birth date, ethnicity, nationality, etc.), their relationships with each other (e.g., parent, spouse, child, etc.), events happened during their life time (birth, marriage, divorce, death, graduation, migration, etc.) related to the individual and/or the family, official documents or other documents generated by or about the family. Information about the physical family records themselves and how to access them should also be collected.

多年來圖書館和檔案館已收集和編目許多家譜。但他們所收集的數據的範圍只限於記錄本身 (源代碼級) 的信息。至於家庭中的個人，只有少數顯著的個人的簡要說明資料記錄在圖書館目錄上。本項目的重點是擴大家譜收集的內容並存儲在一個結構良好的電子格式中。這裡的信息包括個人基本信息 (如姓名，性別，出生日期，民族，國籍等)，他們彼此

之間的關係（例如，父母，配偶，子女等），他們在生期間發生的事件（出生，結婚，離婚，死亡，畢業，遷移等）有關的個人和/或家庭，與該家庭有關的官方或其他文件。家庭的實物記錄以及如何訪問這些記錄的信息也應被收集。

DAZUPU project will be a collaborative project among all interested parties, so it is critical that we agree on what data to collect and how they are recorded.

DAZUPU 項目將是有關各方的合作項目，因此至關重要的是我們大家要同意收集的哪些數據，以及如何記錄這些數據。

1. *System Selection Criterion* 系統選擇的原則

There are many Genealogical Information Management Systems (GIMS) available, some are available as open source software, others are commercial systems. Some vendors or other organizations also provide hosting services.

有許多家譜信息管理系統（GIMS），有些是開源軟件，其他都是商業系統。一些供應商或其他機構還提供託管服務。

Because of the unique nature of this project, we need a system that is not only able to handle data in Chinese (this includes staff and end user functions), but has a complete Chinese user interface.

由於這個項目的獨特性，我們需要一個系統，不僅能夠處理在中國的數據（包括工作人員和最終用戶功能），還有一個完整的中國用戶界面。

The following is a brief list of criterion used to evaluate existing systems.

以下是一個簡短的標準清單，用於評估現有的系統。

General 概論

- Multi-generation pedigree with different views
- 多世代譜系要有不同的視圖

- search by names (first name, last name, both English and Chinese of traditional, simplified and romanized forms)
- 可以按姓名的各種表達方式搜索（名字，姓氏，英文和中文繁體，簡體和羅馬拼音形式）

- Provides other searching methods (events, keyword, birth location, living city, etc.) and ability to set scope of searching
- 提供其他搜索方法（事件，關鍵字，出生地點，居住地等）並可以設置特定的搜索範圍

- Database backend: what backend database server(s) does the system support?

- 數據庫後端：什麼樣的後端數據庫服務器 (S) 系統的支持？
- Data elements: in addition to the standard GEDCOM fields, it should be able to handle additional elements;
- 數據元素：除了標準的 GEDCOM 範疇，它應該能夠處理額外的元素；
- Allows users and site visitors to add additional information, such as accomplishments, pictures, video clips, web site addresses; blog-like feedbacks/comments about the person.
- 允許用戶和網站的訪問者中添加其他信息，如成就，圖片，視頻剪輯，網站地址；博客-如反饋/讀者意見等。
- Online editing: Authorized user should be able to add parent, spouse and children to an existing person; ability to add person and link to an existing person in the tree.
- 在線編輯：被授權用戶能對在譜人物添加父母，配偶和子女；能夠增加人物並鏈接到在譜的人物的家譜樹上。

Printer-friendly output: ability to generate printable report (PDF, XML and/or HTML) based on user-defined criterion. The system should provide some predefined (canned) reports.

- 友好的打印輸出：能夠按用戶定義的標準生成可打印的報告 (PDF , XML 和/或 HTML)。該系統還應提供一些預先設定的報告。
- User authentication/authorization: the system should supports multi-level authorizations, and protect sensitive and secured data.
- 用戶認證 /授權：該系統應支持多級授權，並保護敏感和保密的數據。
- System should be able to expand either Horizontally (expand to multiple Web/Database servers) or Vertically (Upgrade to more powerful Servers).
- 系統應能水平擴大 (擴大到多個 Web 服務器/數據庫服務器) 或垂直擴大 (升級到更強大的服務器)。
- Backup and restore. System should provide the capability of scheduled backups of different types.
- 備份和恢復。系統應提供不同類型的預設備份的能力。

Multi-lingual support 多語種支持

- The system should be able to handle English as well as Chinese (both traditional and simplified). The system should be Unicode-based.
- 該系統應能夠處理英語以及中文 (繁體和簡體)。該系統應基於 Unicode 的。
- User interface should use a Unicode-based encoding, and allows user to select his/her own preferred language (English, Simplified Chinese or Traditional Chinese).
- 用戶界面應該使用一個基於 Unicode 的編碼系統，並允許用戶選擇他/她自己的首選語言 (英文，簡體中文或繁體中文)。

- The search engine should be able to accommodate searching in all three language options (English, Traditional Chinese and Simplified Chinese), as well as searching in romanized Chinese (i.e., pinyin) without the need of inputting data in all three language encoding.
- 搜索引擎應該能夠容納使用所有三種語言（英文，繁體中文和簡體中文），以及漢語拼音。
- Support sorting in Chinese (pinyin and/sort stroke count)
- 在支持中文的排序（按拼音/筆劃數排序）

Data Exchange 數據交換

- Import: ability to import GEDCOM 5.5 file and merge into the system; ability to accommodate future XML formats; ability to import from delimited file;
- 導入：能輸入 GEDCOM5.5 格式文件並併入系統；能適應未來的 XML 格式；能輸入用分隔符處理的文檔文件；
- export: ability to export GEDCOM (the whole tree or a subset of it) and/or XML.
- 出口：能以 GEDCOM 格式輸出（整個家譜樹或它的一個子集）和/或 XML。
- Remote standalone client application: it would be desirable if the system provides a client software that allows user to add records offline and batch-load the records onto the system.
- 能遠程控制獨立客戶端的應用程序：如果系統提供了一個客戶端軟件，允許用戶離線添加到記錄和成批的上載記錄到系統裡。
- Allows authorized user to import family data in GEDCOM 5.5 and/or XML based on GEDCOM, and delimited file online
- 允許授權用戶在線輸入該家庭的數據到 GEDCOM5.5 和/或以 XML 為基礎的 GEDCOM 和分隔文檔文件
- Support data export in GEDCOM or XML based on GEDCOM format via web-based interface. Authorized user should be able select subset of the whole database to export; user should only be permitted to export data that he/she is entitled to.
- 通過互聯網界面，支持數據輸出到 GEDCOM 或以 XML 為基礎的 GEDCOM。授權用戶可以選擇整個數據庫的子集導出，用戶應該只被允許輸出他/她有權限的數據。
- All export should be encoded in Unicode (UTF-8 preferred)
- 所有的輸出應該都是以 Unicode 編碼的文件（UTF -8 的首選編碼）
- Does the system support real time access of data via API or Web Service?
- 該系統是否支持通過 API 或 Web 服務的數據實時訪問？
- Does the system support real time access of data as RSS feed?

- 該系統是否支持 RSS 提要的數據實時訪問？

Next Step 下一步

DAZUPU is a huge undertaking. There is no way that DAZHONGHUA Family Tree Development Association alone will be able to complete this project. Its success relies on partnership with the whole Chinese community worldwide, as well as researchers, librarians and commercial vendors and all other interested parties.

DAZUPU 是一個巨大的承諾。僅靠大中華大族譜協會 (DAZUPU) 是不可能完成的。它的成功有賴於整個華人世界，以及研究人員，圖書館員，商業供應商和其他有關各方的夥伴關係的合作。

A technical team will be formed after the conference to discuss and/or make recommendations on the following issues to the Board of Directors:

在這次會議後的我們將成立一個技術小組，就以下問題討論/研究，並向理事會提出建議：

- Determine the data elements that need to be collected
確定需要收集的數據元素
- Determine the format these data elements need to be in.
確定這些數據元素需要採用的格式
- Determine standard mapping of historical and current names of geographical locations
確定歷史上的和當前的地域名字的標準對應
- Adopt or develop mapping between historical dates and Gregorian calendar dates
採用或開發的歷史日期和公曆日期之間的標準對應 (我們已經完成此項工作)
- Propose a DAZUPU-specific set of GEDCOM data elements (_DZPxxx) to accommodate information that is not currently in GEDCOM specifications.
提出 DAZUPU 特定的 GEDCOM 數據元素集 (_DZPxxx) 的建議，以適應不在目前的 GEDCOM 規範內的信息存儲需要。
- Propose a XML-based specification for data exchange via web service among DAZUPU participating systems
提出基於 XML 的參與 DAZUPU 項目的系統間通過網絡服務進行數據交換的規範
- Examine existing systems (including commercial and community-developed systems) and make recommendations
檢查現有的所有系統 (包括由商業單位和社團開發的系統)，並提出建議