Preface

This February 20th and 21st, a workshop on the ***Possibilities of a Knowledgebase of Tang Civilization -- Towards a new comprehensive digital archive of Tang China*** was held at the Institute for Research in Humanities. As work on the Knowledgebase of Tang Civilazation within the COE project has begun, it seemed appropriate to invite scholars with some experience that could contribute to the development of this project and provide an opportunity for brainstorming and exchange of ideas.

Since work on the digitization of cultural artefacts had already been carried out on a large scale in Taiwan for some years and more recently there had been a systematic effort to coordinate and integrate individual projects on a national level, the decision was made to invite colleagues from Taiwan to introduce us to their views. With good fortune we have been able to assemble a panel that could provide historical, linguistical, geographical and systematical perspectives on the topic and also give examples of different methodological approaches. The papers presented at that workshop are collected here, three of them as formal papers, while the report on Geographical Information Sciences (GIS), which served as a vivid introduction in this field, has been mostly perserved as the original on-screen presentation.

On the day following these presentations, the invited speakers and commentators, as well as staff of the Tang Knowledgebase team and members of the Institute met again in a closed session which lasted the whole day, to discuss the current plan and future possibilities of the project. The day started with a detailed presentation by Christian Wittern about the background, working plan and current state of affairs of the Knowledgebase of Tang Civilization. Prof. Muller as well as the other presenters gave numerous and very valuable comments. The discussions continued after the lunch break with some concrete proposals, for example about how to encode the different calendars that had been used during the Tang dynasty, about how the content of the Knowledgebase could be used to develop a Tang Dictionary. Other topics covered included the establishment of a policy for data sharing amoung involved institutions, as well as implications of the project for education and training of researchers.

The workshop provided ample opportunity for the exchange of ideas and comments and served as a productive correctivum for the plans of the Knowledgebase of Tang Civilization, this is unfortunely only to a very limited extent visible through the papers printed here, but hopefully the work on the Knowledgebase will reflect it more fully. In conclusion, I would like to take this opportunity to thank all participants, the audience and all those who worked hard to make this workshop possible for their nevertiring efforts.

Christian Wittern, June 2004

Contents 目次

文物數位化雜議

On Digitalization of Cultural Collections

謝清俊 Ching-Chun Hsieh
台灣玄奘大學 Hsuan Chuang University, Taiwan, ROC

This paper is dedicated to the memory of the pioneer in digitization, the late Professor Katsumura Tetsuya (1937-2003)
謹以此文記念孤明先發的數位化先驅，勝村哲也教授（***1937-2003***）

Abstract

Nowadays, people begin to realize that any "form" can be digitalized. So, besides digital libraries, there are digital museums and various digital archives. Digitalization becomes so popular that you can find almost anything on the Internet. Digitalization has already begun to influence our daily life, our work, and as a consequence, it will also change our future. With this background, it is very interesting to have this workshop to study the digitalization of Tang Culture. But, there are some fundamental open questions about digitalization remains unanswered. Such as, what is information, what is digitalization, why it is so important, what are the impacts and future perspectives of digitalization on our culture, and what shall we do with digitalization This paper will present some primitive theories trying to address these fundamental questions with the hope that some scholars might be interested in study these topics in more detail. In the beginning, we try to do a review on the relationship among information technology (IT), media, communication and cultural in general. Some philosophical background will also be provided. Then, a general definition of information will be presented. With this definition, a set of very clear concepts of information can be derived. This definition can be served as a consensus for those who participated in the project of constructing a virtual reality for Tang civilization.

In the next topic, we will discuss a model that organizes all kinds of works involved in doing digitalization. This is an empirical model learned from the Digital Museum Project and the National Digital Archives Program in Taiwan some years ago. We like to share this model with you.

資訊、傳播、文化、與媒介

近年來，凡是提到資訊，幾乎人人的第一聯想就是電腦；反之亦然。這實在是一種特奇的現象，批判學派稱之爲一種錯誤的資訊意識型態[1]。其實，資訊和電腦的關係大約只有五十年，從有電腦商品之後才有的。反倒是人們似乎忘記了與資訊唇齒相依的傳播（或溝通）。傳播和資訊關係之密切，可以借用佛典裡的話：「此有故彼有，此無故彼無」[2]來說明。這麼密切的關係，實在不容忽略。

談到傳播，就不能不想到文化。若沒有傳播，生物的社群都不可能出現，更不要說文化了。所以，傳播是文化產生的必要因素。其實，傳播不僅僅是文化的必要因素，更可想像爲文化的基因之一，這是因爲：人類學與社會學作了無數的文化研究，不難歸納出如下的論斷：不同的傳播行爲將導致不一樣的文化[3]。麥克魯漢曾指出的「口語文明」和「文字文明」的差異[4]，這就是個典型的例子。

從上述的資訊—傳播—文化的關係，顯而易見的，資訊和文化的關係也是密切得出乎一般人的想像。如果把文化看作一個系統，無論是自然的還是人工的，那麼，依系統的三要素：物質、能量、和資訊這三者而言[5]，資訊便是文化系統的首要要素。這是因爲，系統的存在、成長、和演進，都依賴資訊的指令的緣故。既然資訊和文化的關係也如此密切，那麼，資訊科技與文化的關係也就不言可喻了。

資訊科技與文化的關係如此密切，也似乎超出常理；然而，這正是可以解釋近二、三十年來社會極速變遷的主因。從另一方面來說，解鈴繫鈴，資訊科技未嘗不是化解兩種文化（the two cultures[6]）現象的橋樑。

然而，資訊科技要作此橋樑，還缺少一個以往被許多學門所忽視的環節，那就是：對媒介材料及其衍生性質的了解。爲簡潔計，在本文中讓我們以「媒介」一詞來表示①媒介材料、②依此媒介材料而創造的工具、③以此工具而衍生的技術、以及④在使用此工具和技術的情境下，所衍生的表現系統（如蘇珊・郎格的符號學美學與文化符號理論，請參考註[8]）。

許多人類學、哲學和文化學者，都同意麥克魯漢將文化進程區別爲：口語文明、文字文明和多媒體文明的劃分。若從媒介的角度觀察，這三個進程與媒介演進的關係，就像是一回事一樣的吻合。所以這三個時期，可以看是媒介轉變而引起的。例如，從有文字以後產生了素養問題（literacy）；要會讀別人寫的作品，要會用文字跟別人溝通。這是因爲媒介改變的緣故。當多媒體文明出現時，也產生了素養

---

[1] Jennifer Daryl Slack and Fred Fejes ed., **The Ideology of the Information Age**, Ablex, 1987

[2] 《中論》龍樹菩薩造。

[3] James W. Carey, **Communication as Culture**, Unwin Hyman, 1989

[4] 此處並未直接引用麥克魯漢所用的名詞，僅依劃分文明的內容說明。請參照 Marshall McLuhan, **Understanding Media**, McGraw-Hill, 1964

[5] 這是依據 Norbert Weiner 的系統理論。

[6] C. P. Snow, **The Two Cultures**, Cambridge University Press, 1959.

問題，那就是近三十年來陸續出現的電腦素養、資訊素養、網路素養…等，而素養造成的社會問題便稱爲數位落差（digital divide）。

從口語到文字媒介的轉變，導致文字文明。這是眾所周知的事。依此模式，要檢視從文字文明進入多媒體文明的媒介轉變，就必然要了解數位媒介。在文字文明時期，媒介材料是物質。任何記載都依賴物質，而用過的媒材便不可再用（加工後已不是原來的媒材）。所以，文字文明可說是物質媒材的文明；此文明的任何記錄，無論是知性的還是感性的，都必須受制於當時的經濟法則（因媒材爲物質故），也都免不了有物質障礙，如有重量、佔空間、會破損、失竊等。

數位媒介和物質媒介有兩個基本的差異。其一是，數位媒介是以能（energy）爲媒材。較精確的說法是：它利用物質內部穩定的能階（energy state）狀態來儲存符碼，所以也稱之爲能階媒介。能與物質的性質有極大的差異，數化的虛擬世界之所以能夠超越物質時空、超越物質障礙，就是這個原因。文物數化後，搖身一變，變成幾乎取之不盡、用之不竭的資源，實源於此。

其次的差異是，數位媒介是唯一的媒介，是唯我獨尊的。在物質媒介時期，有各式各樣的媒材，如竹、木、泥、石、帛、紙等。這是多樣媒介共存的時期。然而，到了多媒體時期，什麼媒材都可轉化成數位形式；於是，數位媒介號稱多媒體，因爲它取代、統一了所有的物質媒材，唯我獨尊。

這唯我獨尊的情勢，造成了完全出乎想像的後果。比方說，任何機器都可合併，如電話、傳真、影印、答錄機、時鐘、收音機、計算器…可以做成一個。又如，不僅僅是語文（即口語至文字文明的媒介轉變），連所有的物質形式，都可以數位化。這就是爲什麼「文、物」皆可數化的由來。

從口語轉變至文字文明的過程中，所有文化傳承的記錄，均須由口耳相傳的方式，轉變爲文字的表達。例如，佛陀說法四十九年，之後約經歷五百年才將之轉化爲文字記錄。如果沒有這轉化，今日是否有佛經誠屬疑問。從文字轉變爲數位能階媒介，也如同上例，關係到今後文化的承先啓後，甚至於文化的存亡絕續。是故文物的數位化，對文化而言也有無比的意義。

資訊的界說

資訊的界說很有趣。最近有些文章裡說到：從 1955 年到目前，什麼是 Information 一直困擾著我們，至今沒有一個學科或行業可以把什麼是 Information 搞清楚[7]。處此情境，要討論唐文化的數位化或文物的數位化時，這個陰魂不散的問題又出現了；試想，若是大家對 Information 的概念、想法都不一致、沒有共識的話，怎麼來做文物的數位化呢？所以，我們很需要一個簡單而通用的

[7] 請參考 A.D. Madden, **A Definition of Information**, Aslib Proceedings vol. 52, No.9, p.343-, 2000 或：牛津大學 Lucino Floridi 教授今年四月將在 ***Metaphilosophy*** 雜誌上發表的論文，和在 ***Minds and Machines*** 雜誌主編的，有關資訊哲學（philosophy of information）的一些論文。

information 的定義，作爲討論、研究和實施數位化的前題。

讓我們以蘇珊・郎格(Susanne K. Langer, 1895-1982)的符號學的美學 (或稱符號論美學)作爲引子。符號論美學在二十世紀的五十年代主導美國的美學思潮，它把文化看成由符號所組成，在人文主義跟科學主義方面都有較大的包容性。[8]文化符號可分爲兩類：一類是推理符號（是內涵概括確定的理性符號，它既可以翻譯，也可以被分解、推理；如語言符號。）一類是表象符號（是非理性的、完整獨特不能被分解的，具豐富含義的情感意象；如藝術符號）。藝術可作爲一種傳播的文化符號，文學、歷史也可以這樣看。蘇珊・郎格說：「藝術是表現人類情感的符號形式」。

文化是什麼？人們概略地說文化是人們生活的表現。卡西爾(Ernst Cassirer)說，文化是人類所知客觀化的過程[9]。卡西爾界說對文物數話而言，是很適合的。然而，無論怎麼說，所知是一切人類行爲的根本動力；藝術是依據所知的感性面而爲之，生活是常識面，學問是理性面。所以我們可以說，人們的社群累積了什麼樣的所知，就形成什麼樣的文化。是故我們可以採用：資訊就是『**所知表現在媒介上的形式**』的界說[10]，來建立文物數位化的共識。這兒的「所知」指的是心智上的一些活動，包括理性的、感性的、信仰的、意志的都在內。據此定義，資訊是實際存在的形式(form)，所以它可以數位化，電腦可以處理它。資訊是屬物性的，而所知則是心智的。資訊承載著所知；資訊是所知的形式，所知是資訊的內容。由所知到資訊是外化的過程，也是從心智轉化爲實物的過程。

「資訊即**所知表現在媒介上的形式**」是依據一般資訊產生的過程界定的，包含資訊生起的主要因、緣、果；這產生的過程本身就是一傳播行爲。依此定義，資訊的性質可以劃分爲四類，即：①因襲所知的性質，②繼承媒材的性質，③依媒介工具和技術所增益的性質，和④表現系統所呈現的性質。每一類之下，還可以列出許多較細的項目，請參考表一。

我們都曉得電腦的能力強大，但經常感到困惑的問題是：電腦能作什麼事？不能作什麼事？這個問題如果問電腦專家，他會告訴你：「問題的解法如果複雜到某某程度以上，計算機就沒辦法作。」這是從問題的複雜程度來看的。如果從美學或哲學的角度來看的話，其實非常單純：「電腦只能處理形式，它不會直接處理問題的內容」。這裡所說的形式及內容都是美學裡面的辭彙。因爲電腦可以直接處理形式，所以任何美學、文學，只要把它看成一種符號，無論是文化符號、推理符號、表象符號，不管它是文字還是色彩，只要是一種形式，電腦就可以設法處理。這是爲什麼電腦可以踏進文學、美學

[8] 對此有興趣的讀者請參考唐孝祥、袁忠、唐更華編著之《美學基礎教程》華南理工大學出版，1998 初版（2002 重印），第四章西方美學發展歷程，第 149 頁。

[9] Ernst Cassirer 原著，于曉等譯《語言與神話（Sprache und Mythas）》桂冠文化，1990

[10] 講者已使用此定義多年，請參照〈**談資訊的定義與性質**〉謝清俊, 「資訊科技與社會轉型學術研討會」中研院社會學籌備處, Dec. 1996

或文化領域做文物數化的基本理由。

表一、數位資訊的性質

壹、 因襲所知的性質

一、 所知是心智的能量所現，它指導一切行爲，包括理性的、感性的、創造力、意志力等。
二、 所知無所不在，資訊亦然。
三、 所知可以匯集、增長，資訊亦然。資訊的匯集將產生綜效（synergy）。
四、 資訊是一個系統中最主要的資源，它管理著系統中的物質和能量，是系統能存在、成長、演進的基本推動力量。

貳、 依附數位能階媒介所得之性質

一、 能階媒介擺脫了物質障礙，時空障礙極低，有取之不盡、用之不竭的效果。
二、 將能階媒介數位化的結果，產生了數位能階媒介。任何傳統所用過的物質媒介，均可轉換爲數位能階媒介。於是，數位能階媒介成了唯我獨尊的媒介，而各種設備間的互通性、相融性大爲增加，使設備間可以溝通，甚至於合併。
三、 數位能階媒介是傳播和資訊的基因，凡是用到數位能階媒介的傳播或資訊皆有以上的性質。

參、 駕馭媒介工具與技術所增益的性質

一、 電腦處理資訊的能力
二、 資訊基礎建設（information infrastructure）所提供的方便環境。
三、 軟、硬體的交相配合、替代
四、 各種介面規範的相容和劃一，成爲重要趨勢和議題。
五、 以機器駕馭所知（知識）
   1. 資訊的匯集經增加彼此參照的結果，產生「總體大於部份之和」的效果。
   2. 數位化導致一切既有所知的重新整理；於此也創造了嶄新的工作環境。
   3. 數位化的綜效改變了行業間既有的依存關係。
   4. 提供解決複雜社會問題的新希望。

肆、 表現系統改變所呈現的性質

一、 多媒體的普及改變了傳統的語文的現象。
二、 傳播的觀念、方式、程序、效果皆產生變化。傳播的變化，改變了社會的依存關係，人際關係與生活的方式，如食、衣、住、行、育、樂等。

三、 文物的數位化涉及如何將文物相關的所知，表達在電腦中的問題。此即一典型的知識表達問題（knowledge representation problem），爲解決此問題，標誌系統應運而生。標誌系統的應用，意味著人工智能和文獻學的結合，此結合形成人類所知一個全新的記錄系統。

文物數位化之層次

一個文物的數化可以分三個層次：

其一是**外觀的數位化**：這部分大家最喜歡看、最知道怎麼做、問題也最少。通常是把文物掃描，如文獻中的文字、版面、照片...等。一件文物的數位化，外觀是必要的基礎成份。

其次是**背景資料的數位化：**舉凡一份文件的書目、作者、譯者、屬性、圖說…等都屬此類。背景資料是一些客觀的資料，包含有時下流行的後設資料（metadata，或譯爲元資料）。背景資料的對錯是只可以考證而不容詮釋的。

再次是**內容的數位化：**內容的數位化包含對內容的注解、詮釋、分析、考據等，目前所做的還不是很多。內容的詮注可以依某一個文獻、主義、理論、研究的方法…等作爲立論所在。內容的數位化即是文物相關知識的整理和再現，是最重要的；沒能做到這一部份，數化就失去了應有的深度和廣度。

以上只說明了一件文物數化的內涵。文物之間是有密切關聯的，這部份就涉及到文物整體數化的問題了，也就是文物數化的第四層次—**相關資料的參照**。相關資料的參照可以分三部分：

互爲文本 (Inter-textuality)：***Julia Kristeva*** 說，任何文學史學之間沒有獨立存在的文字，這些文字一定跟其他的資料之間有某種關聯[A❶]。如果把這些關聯找出來就是一個文學的知識庫（knowledge base）。文史資料跟其他各學科之間還有關係，這類不限領域文件之間關係都屬互爲文本的參照。

情境(context)的參照：從背景說，情境的參照可分爲文化背景、社會背景、以及個人背景的情境；從傳播的角度來講，有作者情境及讀者情境[A❷]。早期的口語傳播沒有這個問題，因爲口語傳播是面對

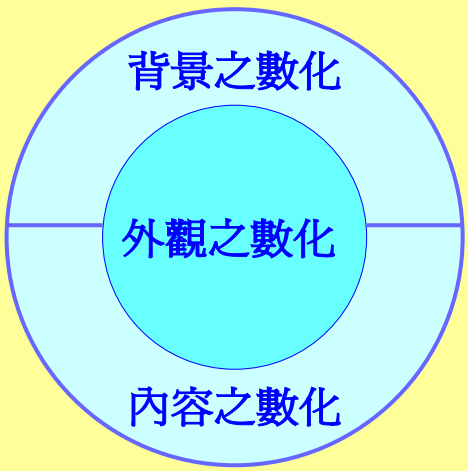
A❶ 關於此說，建議可以參考葉嘉瑩著《清詞選講》三民書局，1996.08，第 115 頁中所說及其例。

A❷ 關於傳播學和資訊學對作者情境及讀者情境的觀點，可參考❷。關於詮釋學對情境的觀點，可參考葉嘉瑩著《詞學新詮》桂冠，2002.02，第 37 頁至 42 頁中所作之說明及其例。

面，讀者跟作者處在同一個情境。當傳播或溝通的機制越來越進步，便幾乎把作者情境及讀者情境斬斷，成爲越來越獨立而交集越來越少的狀態，越分越遠。情境的參照是目前數化尚未接觸的領域，這些不同情境的資料怎麼表達，是文物數位化面臨到較大的挑戰。

文物彼此的參照：如紅樓夢一書中對食譜、建築、花卉…等的參照就是很好的例子。這是人文跟自然的結合。

依作者的淺見，如果本文中的論點無誤的話，那麼，文物數位化的工作將是①對傳統知識和物品的再整理，②是把傳統上文字和器物所承載的所知，轉換到數位能階媒介上的巨大工程，也是③以多媒體虛擬實境再現過去的一個過程，而文物數位化是勢所難免、勢在必行的千秋之事，不是一時之事。如今應該正是文物數位化時機。



圖一：文物數化的層次

結語

本文試圖提出一些觀點，從理論上來了解文物數位化的緣起、性質、趨勢、以及數位化可能對目前人文、社會的影響。

新聞與歷史—

中文新聞內容標誌研究對建構唐代文明知識庫的意涵

News & History：the implication of a study of Chinese news content markup in construction a knowledgebase of Tang civilization

謝瀛春　Ying-chun Hsieh
台灣國立政治大學 National Chengchi University, Taiwan R.O.C.

Abstract

This paper is focused on discussing the core theme of news, i.e., including the six elements of an event what contained some values of the nature of news, also the advent of news in human history. Meanwhile, the Chinese news content markup using XML and the implication of news events study for historical studies are included.

From the history of Journalism, news is highly related to the human history. According to Wang Hung-chuen's book 新聞採訪學 [News Reporting], the early collections of ancient poems in Chinese are viewed as the very early beginning form of news. In today's term, those poems (詩歌) and Chun Chiu (春秋) are the first form of story-telling, i.e., the nature of news. In Rome, Julius Caesar's (100-44 B.C.) daily records are told as the earliest form of news. The nearest modern news form in Europe in the 15th century is the widely distributed "news" (so-called as newsletter nowadays).

"Today's news is Tomorrow's history" is the first lesson taught in journalism school. It is true in deed; news in newspapers and other mass media is always a reference library for historians. In this sense, the analysis of news event using markup languages like XML is a reference for the digitalization of ancient historical archives in the future.

前言：新聞和歷史的關係

新聞學開宗明義指出：今日的新聞即明日的歷史。誠然如此。新聞即翔實記錄每日發生之國際要聞、國家大事、社會重要事件。新聞既是當代重要之記實，因而成爲未來研究當代歷史之重要文獻。從新聞學書籍中可見一斑。

董橋（1997，頁 40～43）論及鄧小平之悼念儀式新聞時，引伸其對歷史人物在史家的評價下難免偏頗。而歷史要寫得好，必須如清代章實齋所說：「譬之人身，事者其骨，文者其膚，義者其精神也」，亦即史實、文筆和觀點三者相互參照。他更進一步說（1997，頁 43），「新聞永遠是歷史的初稿，也是歷史的源頭」，新聞記者應秉持「記」而不「議」的原則，據實報導，當可成爲明天的《日知錄》（按：清朝顧炎武之作）。

馬星野（1970，頁 61-69）論及新聞、歷史和速記的關係時，亦明確指稱：新聞記者與速記人員是不可分的；二者就如古代之左史右史之史官。漢書藝文志：「左史記言，右史記事，事爲春秋，言爲尚書」。

新聞教學即秉持「新聞如記史」之理念爲之。縱使有諸多新聞報導不符此理想，但本此原則者仍居新聞界多數，而純淨新聞之報導亦是以事實爲主。

最早的新聞(東西方之例)：新聞是歷史的初稿

王洪鈞（1991，序，頁 5）指出，中國最早的詩歌及春秋，應視爲人類傳播事業之開端。根據朱傳譽的研究（1967，頁 1-8），《春秋》類似新聞紀錄，而民間的謠諺有如今之「短評」、輿論，是多數人意見的反映。錢震（1967）則認爲新聞一詞是在宋朝已出現，只是其意涵不及今日之「新聞」嚴謹；「是未成熟而又不太可靠的報導」（頁 27）。

在歐洲，羅馬凱撒（J. Caesar, 100-44 B.C.）大帝的每日紀聞爲新聞源頭，而十五世紀時流行於歐陸之新聞信爲近代報紙之雛形（王洪鈞，1991，序，頁 5）。依據小野秀雄著作《內外新聞小史》（陳固亭譯，1964，頁 2-3），羅馬的 Acta Diurna 類似今之報紙，是報紙的始祖。而中國唐朝的邸報和宋朝的朝報，可說是世界上最早的官報和定期刊物。這些都是人類新聞的歷史。

由上述二例可知，歷史與新聞之密切關係。因此，可以說研究新聞的內容及其寫作結構，當可供分析歷史研究之參考。本研究報告在此次研討會之意義即此。

新聞是事件：新聞事件、新聞要素構成新聞內容

新聞必須有事件(event)，並包含六要素，即人(who)、事(what)、時(when)、地(where)、爲何(why)、如何(how)等基本要素，如此才構成新聞內容。

要言之，根據舒曼(E. L. Shuman, 1894)的解釋(Frank Luther Mott, 1952, p.158)，新聞六要素中，人是指新聞事件之主角(subject)，可以是人、動物、機構等；事是指已發生、正在發生或將要發生之事情；時是指事件發生之時間；地是指事件發生之地點；爲何是指事件發生之原因；如何則是指事件發生之情境、過程。

而根據傳播及媒介研究字典(1997, pp. 78-79)，一個事件必須符合媒體採訪報導的新聞價值判準，才可能成爲新聞事件(news event)。新聞事件又可分爲關鍵事件(key event)、類似事件(similar event)及主題相關事件(thematically related event)。

此外，根據新聞實務，新聞事件實際上又可細分爲主要事件、次要事件、其他事件、背景說明，以及細節等類型。本研究即以此爲事件標誌的依據。而新聞寫作的標誌，則根據新聞寫作結構原則，分爲導言(lead)、主體(body)、結尾(ending)分別標誌。

中文新聞內容標誌研究之目的

本研究之目的有六：

1.建立中文新聞內容正式通用的標誌方式

2.提供未來中文新聞內容資訊交換之參照、依據

3.提供新聞寫作教學(像如何寫新聞)之輔助工具

4.提供新聞界修改、編輯新聞(像新聞六要素是否遺漏)之檢查工具

5.使用者可以深入檢索新聞內容(像新聞事件之關係)，不受全文檢索關鍵字詞之限

6.研究用途(像內容分析，探究新聞事件之開始、過程至結束)

研究分析程序：僅以科學新聞之例說明

分析程序如下：20 則科學新聞→DTD→標誌內容→校正(validation)

從 DTD 至標誌內容是以 XML 標誌，從標誌內容至校正間之 SP 是以 Ultra Edit Textpad 工具校正。(有關此研究之中文科學新聞的 DTD，詳見附錄一)

先分析中文新聞內容，以科學新聞，且限於純淨新聞(不包括特寫、 專題報導、深度報導、調查報導及評論)，再以 XML 標誌中文新聞內容。標誌的內容包括標題、消息來源、新聞內容(導言、主體、結尾)、新聞故事(新聞事件、人、事、地、時、爲何、如何)。標誌時以語意單位爲標誌區分之依據；標誌之基本單位是以新聞內容前後文關係之語意完整爲準，而非以句子之文法結構爲依據。

中文新聞內容標誌之例：科學新聞、數位典藏通訊

本研究於 2000 年開始，爲嘗試研究階段，曾中斷一年、斷斷續續進行。最初以科學新聞爲實驗案例，分析了二十則中文科學新聞，且限於純淨新聞(straight news)；因案例最符合制式新聞寫作的要求，而純淨新聞也是最普遍通用的寫作模式。本研究依據新聞寫作之學理與實務，並參照文件製碼協定(TEI; Text Encoding Initiative)及後設資料(Metadata)，圖書館界都柏林核心欄位協定(Dublin Core)及報業資訊交換格式協定(NITF)等進行內容分析、標誌。詳見附錄二中文新聞內容標誌實例。

目前(2004)本研究嘗試以數位典藏

通訊電子報(NDAP Newsletter; the Newsletter for the National Digital Archives Program)內容爲實驗測試對象，但人力(缺乏懂 XML 及新聞之雙重專長者)是最大困難。數位典藏通訊之內容標誌實例詳見附錄三。

結論：中文新聞內容標誌對唐知識庫建構的意涵

中文新聞內容標誌之研究，從新聞六要素；人、事、時、地、爲何、如何，乃至新聞事件，以及新聞要素與新聞事件之彼此關係等研究分析，並以 XML 標誌，成爲電腦可以辨識處理之數位化資料。此研究經驗完全可以轉換至歷史研究。

史學研究對歷史記載、歷史事實、歷史事件、歷史人物、歷史王朝、歷史古蹟，以及諸侯邦國等所有歷史有關的內容，都需要辨認其人、事、時、地、爲何、如何等基本要素。而歷史中各事件間及其人、事、時、地、爲何、如何的關係亦是必須釐清的。這正是歷史內容數位化及其內容標誌之基本工作，也可供唐代文明知識庫建構之參考。

唐代文明知識庫之建構，當作爲未來研究教學，以及查詢、檢索之重要資源。將來勢必考量使用者的基本需求；亦即需要知道歷史記錄內容之人、事、時、地、爲何、如何、各事件及其間的關係。

此外，中文新聞內容標誌之研究，亦考量到以中文語意(semantic)單位爲標誌之基礎單位，而非以語法(syntactic)分析爲標誌依據。因此，此經驗亦可供以中文內容爲大宗的唐代文明知識庫建構之參考。

同時，另一可供參考的則是：中文新聞內容標誌之研究發現，跨學門、跨領域(如資訊科技、新聞、語言、寫作、標準、管理等)的團隊合作(collaboration)，以及合作研究人員之傳播溝通能力及移情心理(empathy)等都是數位化研究工作成敗關鍵(Hsieh, et al, 2003:268-285)。

參考書目

1. 王洪鈞編著，新聞採訪學，台北：正中書局，1991 年。
2. 朱傳譽著，宋代新聞史，台北：中國學術著作獎助委員會，1967 年。
3. 馬星野著，新聞與時代，台北：雲天出版社，1970 年。
4. 陳固亭譯，小野秀雄著，各國報業簡史，台北：正中書局，1964 年（台再版）、1959 年（台初版）。
5. 董橋，新聞是歷史的初稿，香港：明窗出版社，1997 年。
6. 錢震著，新聞論（上），台北：中央日報社，1967 年。
7. A Dictionary of Communication and Media Studies, Arnold (Fourth edition), 1997, pp.78-79
8. Ying-chun Hsieh, Ching-chun Hsieh & John A. Lehman, “Chinese Ethics in Communication, Collaboration, and Digitalization in the Digital Age,” Journal of Mass Media Ethics, 18(3&4), 268-285, 2003.
9. Frank Luther Mott, News in America, Mass.: Harvard University Press, 1952, p.158

附錄

附錄一、中文科學新聞 DTD

```
<?xml version="1.0" encoding="big5"?>
<!--NEWS XML encoding date: 2000-1-4-->
<!DOCTYPE 科學新聞 [
<!ELEMENT 科學新聞      (#PCDATA |科學新聞內容)*  >
<!--                    寫作結構                              -->
<!ELEMENT 科學新聞內容 ( 標題 | 來源 | 導言| 主體 | 結尾)*>
<!ELEMENT 標題          (#PCDATA)>
<!ELEMENT 來源          (#PCDATA|記者名|報紙名|通訊社名|外電日期|刊載日
期|發稿地點|
                                    版面位置 )*>
<!ELEMENT 記者名        (#PCDATA)>
<!ELEMENT 報紙名        (#PCDATA)>
<!ELEMENT 通訊社名      (#PCDATA)>
<!ELEMENT 外電日期      (#PCDATA)>
<!ELEMENT 刊載日期      (#PCDATA)>
<!ELEMENT 發稿地點      (#PCDATA)>
<!ELEMENT 版面位置      (#PCDATA)>
<!ELEMENT 導言          (#PCDATA |事件)*>
<!ELEMENT 事件          (#PCDATA |人|事|時|地|如何|爲何)*>
<!ELEMENT 人            (#PCDATA)>
<!ELEMENT 時            (#PCDATA)>
<!ELEMENT 地            (#PCDATA)>
<!ELEMENT 事            (#PCDATA |人|時|地|如何|爲何)*>
<!ELEMENT 如何          (#PCDATA |人|事|時|地)*>
<!ELEMENT 爲何          (#PCDATA |人|事|時|地)*>
<!ELEMENT 主體          (#PCDATA |事件)*>
<!ELEMENT 結尾          (#PCDATA |事件)*>
<!ATTLIST 科學新聞內容    id     ID    #REQUIRED>
<!ATTLIST 事件            id     ID    #REQUIRED
          類型       (主要|次要|其他|背景說明|細節)    " 細節 "
陳述方式  ( 事實|評論|夾敘夾議 )   "事實"
    內容性質  (新聞消息|科學消息 )"新聞消息"
相關事件  CDATA #IMPLIED
```

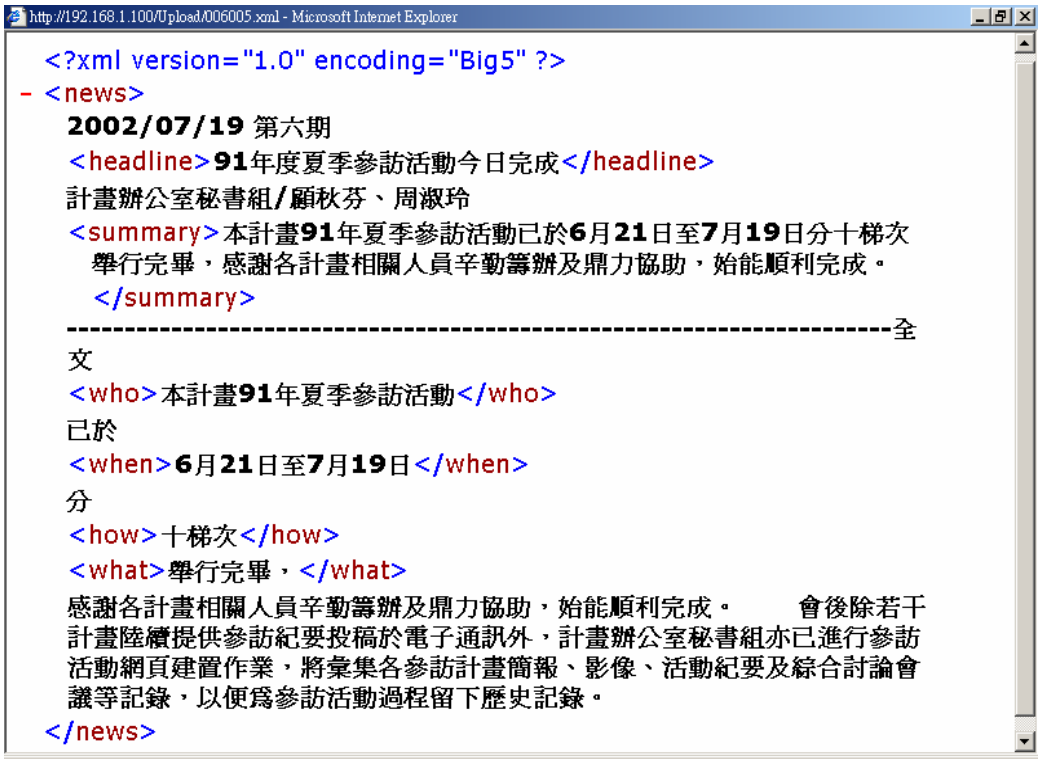
```
            關係類型  CDATA #IMPLIED>
   <!ATTLIST 人        id    ID   #REQUIRED>
   <!ATTLIST 事        id    ID   #REQUIRED>
   <!ATTLIST 時        id    ID   #REQUIRED>
   <!ATTLIST 地        id    ID   #REQUIRED>
   <!ATTLIST 如何      id    ID   #REQUIRED>
   <!ATTLIST 爲何      id    ID   #REQUIRED>
   <!--End of NEWS DTD-->
    ]>
```

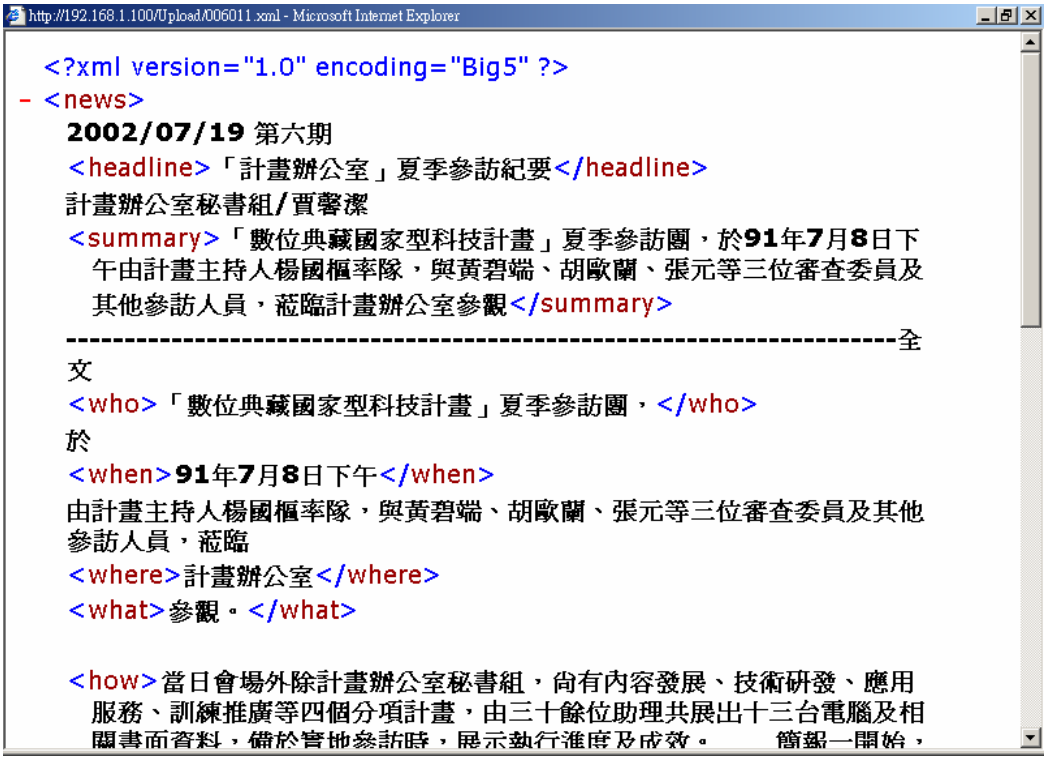
附錄二、中文科學新聞內容標誌實例

```
  <?xml version="1.0" encoding="big5" ?>
- <!-- NEWS XML encoding date: 2000-1-4-->
 <!DOCTYPE 科學新聞 (View Source for full doctype...)>
- <科學新聞>
- <科學新聞內容 id="SN2">
<標題 />
  - <來源>
   <報紙名>中央日報</報紙名>【
   <發稿地點>中央豪士敦</發稿地點>
   <外電日期>五日</外電日期>
   <通訊社名>美聯</通訊社名>電】</來源>
- <導言>
- <事件 id="M1" 類型="主要" 陳述方式="事實" 內容性質="新聞消息">
  <人 id="A1">第一位接受純粹人工心臟移植人</人>，
  <時 id="C1">今天</時>
  <事 id="B1">清醒著，情況令人滿意</事>。
  <人 id="A2">他的妻子</人>
  <事 id="B2">則淚汪汪的懇求有人捐贈人類心臟</事>。</事件> </導言>
- <主體>
- <事件 id="De1" 相關事件="M1" 關係類型="補充" 類型="細節" 陳述方式="事實"
  內容性質="新聞消息">
  <人 id="A3">伊利諾州人卡普</人>，
  <時 id="C2">前天</時>
  <事 id="B3">在三小時的手術中，接受移植一個實驗性的人工心臟</事>。
  <事 id="B4">他的心室嚴重受損，修補無望</事>。</事件>
- <事件 id="De2" 相關事件=" De1" 關係類型="補充" 類型="細節" 陳述方式="事實
```

" 內容性質="新聞消息">
<人 id="A4">**曾動過十八次心臟移植手術的聖洛克主教派教會醫院心臟移植小組主持人庫里醫生**</人>
<事 id="B5">**說，純粹機械的心臟，過去僅用於動物**</事>，
- <爲何 id="F1">**而**
<事 id="B6">**這次手術，則僅用以維持病人活著**</事>，
<事 id="B7">**等待有人捐贈心臟**</事>。</爲何> </事件> </主體>
- <結尾>
- <事件 id="De3" 相關事件="M1" 關係類型="補充" 類型="細節" 陳述方式="事實" 內容性質="新聞消息">
<人 id="A5">**醫院發言人**</人>**在**
<人 id="A6">**卡普的妻子雪莉**</人>
<事 id="B8">**懇求捐贈心臟後不久說**</事>：「
<事 id="B9">**目前我們最關心的是一個合適的捐贈人**</事>。」</事件> </結尾>
</科學新聞內容>
</科學新聞>

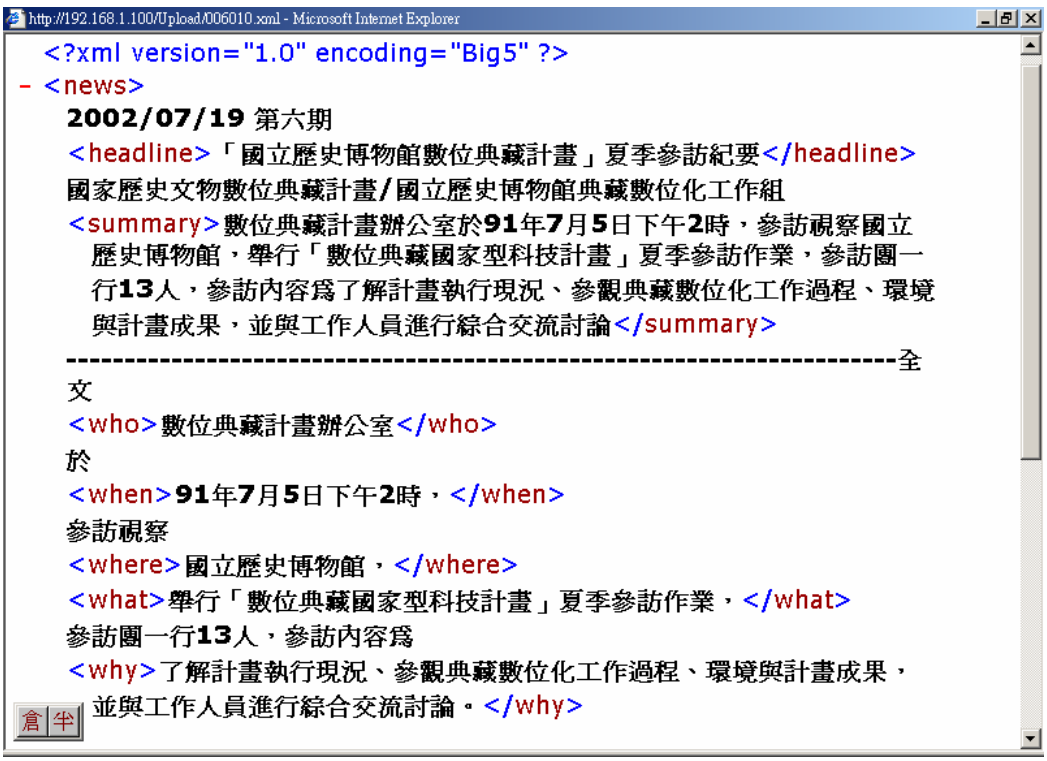
附錄三、數位典藏通訊之內容標誌實例




http://192.168.1.100/Upload/006011.xml - Microsoft Internet Explorer

<?xml version="1.0" encoding="Big5" ?>
- <news>
2002/07/19 第六期
<headline>「計畫辦公室」夏季參訪紀要</headline>
計畫辦公室秘書組/賈馨潔
<summary>「數位典藏國家型科技計畫」夏季參訪團，於**91**年**7**月**8**日下午由計畫主持人楊國樞率隊，與黃碧端、胡歐蘭、張元等三位審查委員及其他參訪人員，蒞臨計畫辦公室參觀</summary>
----------全文
<who>「數位典藏國家型科技計畫」夏季參訪團，</who>
於
<when>**91**年**7**月**8**日下午</when>
由計畫主持人楊國樞率隊，與黃碧端、胡歐蘭、張元等三位審查委員及其他參訪人員，蒞臨
<where>計畫辦公室</where>
<what>參觀。</what>

<how>當日會場外除計畫辦公室秘書組，尚有內容發展、技術研發、應用服務、訓練推廣等四個分項計畫，由三十餘位助理共展出十三台電腦及相關書面資料，備於實地參訪時，展示執行進度及成效。　　簡報一開始，


http://192.168.1.100/Upload/006010.xml - Microsoft Internet Explorer

<?xml version="1.0" encoding="Big5" ?>
- <news>
2002/07/19 第六期
<headline>「國立歷史博物館數位典藏計畫」夏季參訪紀要</headline>
國家歷史文物數位典藏計畫/國立歷史博物館典藏數位化工作組
<summary>數位典藏計畫辦公室於**91**年**7**月**5**日下午**2**時，參訪視察國立歷史博物館，舉行「數位典藏國家型科技計畫」夏季參訪作業，參訪團一行**13**人，參訪內容爲了解計畫執行現況、參觀典藏數位化工作過程、環境與計畫成果，並與工作人員進行綜合交流討論</summary>
----------全文
<who>數位典藏計畫辦公室</who>
於
<when>**91**年**7**月**5**日下午**2**時，</when>
參訪視察
<where>國立歷史博物館，</where>
<what>舉行「數位典藏國家型科技計畫」夏季參訪作業，</what>
參訪團一行**13**人，參訪內容爲
<why>了解計畫執行現況、參觀典藏數位化工作過程、環境與計畫成果，並與工作人員進行綜合交流討論。</why>

Two Historic GIS and Their Applications on Electronic Databases

范毅軍 I-Chun Fan, 廖泫銘 Hsiung-Ming Liao

台灣中央研究院 Academia Sinica, Taiwan

Abstract

Space and time are the two most important elements that constitute changes in history, culture, and environment. Through computer technology, traditional maps can be transformed into virtual reality which can be updated and integrated with various database. Before our projects, there was no adequate digital Chinese and Taiwanese history, culture and natural resources atlas which can serve as a comprehensive reference for scholars in Sinology studies. The CCTS Project (Chinese Civilization in Time and Space) which began in 1996 aims to build up a fundamental system for historical GIS and basic data of Chinese history of the past 2000 years. The THCTS Project (Taiwan History and Culture in Time and Space) which began in 2001 aims to develop a spatial-temporal application infrastructure on the basis of digitized Taiwanese history, culture and natural resource maps of the past 400 years.

Both of these two systems consist of three components: geospatial base data, thematic map database, and WebGIS-based application. Take THCTS for example. The development of Taiwanese history can be divided into Dutch and Spanish Period, Koxinga Period, Qing Dynasty, Japanese Colonial Period and After WW II. Accordingly, we have developed various base maps of each period. Among them, the *Atlas of the historic administrative pau division of Taiwan* (1904), *Atlas of the topography of Taiwan during the period of Japanese rule* (1920) and the newly published *Topography of Taiwan* constitute the main body of the base maps.

On the basis of these geospatial base maps, we have integrated available researches in Taiwanese history and culture. Nine groups of thematic maps were produced, such as population distribution, religion, education and aboriginal people. In addition, through Internet and GIS functions, we integrated various Taiwanese maps and important research projects, such as "Taiwan Studies over the Internet", "Taiwan Aerial Photo Management System" and "Taiwan Gazetteer". With this integration, we developed interface for data search and application, and further facilitated ongoing research projects of different kinds.

In addition to providing digital geospatial and thematic maps, there are other applications of the two systems, CCTS and THCTS. Take CCTS for example, the first application is the integration and retrieval of electronic database. The CCTS system serves as an integrated tool for searching the "Scripta Sinica" and the bibliography of Chinese Local Gazetteers. One can use this system to easily search the "Scripta Sinica" and the bibliography of Chinese Local Gazetteers through keywords or directly through the CCTS Web-GIS system, which allows users to locate a certain place or geographic area on the digital map and search for Chinese local histories related to that area. The second application of CCTS is the application of digital archive projects. The CCTS system is in collaboration with such digital archive projects as "Distribution of the Han Tombs and Temples" and "Mapping the Journeys of the Song-dynasty Poet Su Dongpo." The third application of CCTS is spatial analysis. On the basis of various geospatial base data, researchers and users can create their own maps, integrate with available information and conduct spatial analysis using the CCTS system. Research projects such as "Overlay Analysis of Ming-Qing Jiangnan Market Towns" and "Buffer Analysis for the Flooding of the Yellow River during the Tang Dynasty" are examples.

The CCTS and THCTS are Web-GIS based infrastructure for content navigation and web mapping. These two systems integrate a variety of geospatial information and digital database via the Internet. In the future, we hope to make more value of this spatio-temporal infrastructure through active participation from scholars across different fields.

Chinese and Taiwan Historic GIS

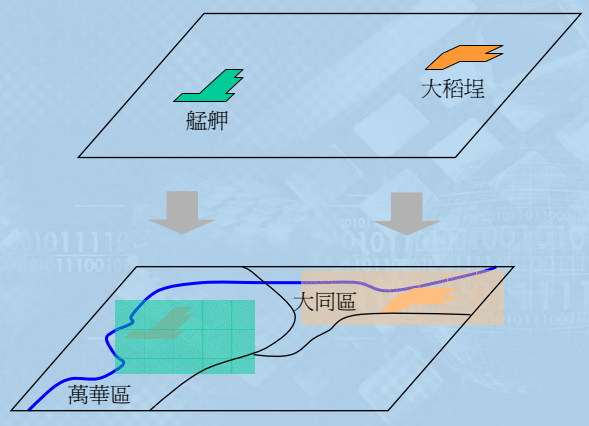
(Chinese and Taiwan Civilization in Time and Space)

Fan, I-chun 范毅軍(中央研究院歷史語言所副研究員)
Liao,Hsiung-Ming 廖泫銘(中央研究院計算中心組長)

Table of Contents

Project Background and Objectives

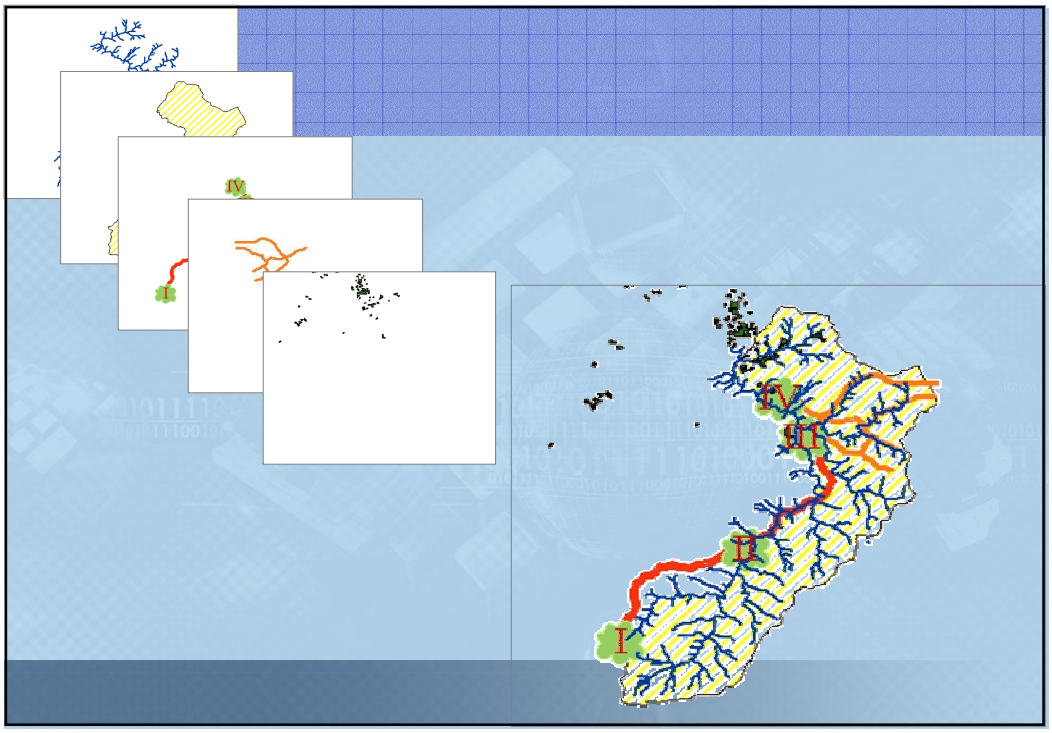
- Space and time are the two most important elements that constitute changes in history, culture, and environment.
- Before our projects, there was no adequate digital Chinese and Taiwanese history, culture and natural resources atlas which can serve as a comprehensive reference for scholars in Sinology studies.
- Through computer technology, traditional maps can be transformed into virtual reality which can be updated and integrated with various database.

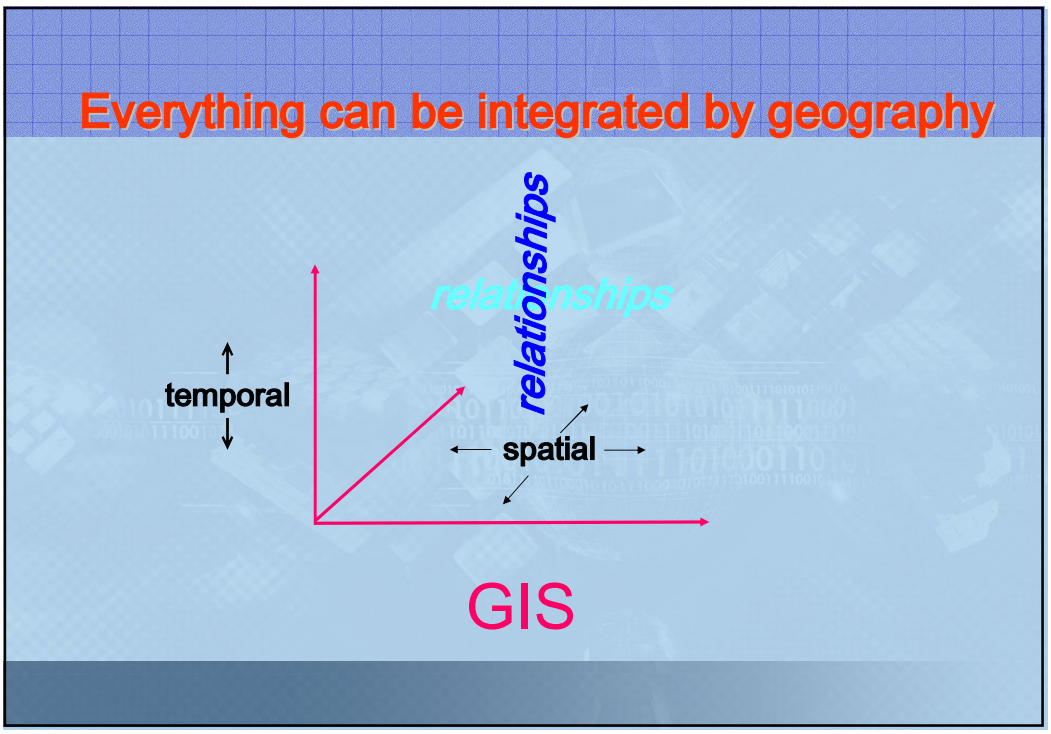


part of
close to
near
adjacency
used to be
....

空間化 → 絕對位置化 → 整合 → 空間關係推導

Explicit Spatialization → Absolutely positioning → Integration →Spatial Analysis


IV
I
IV
III
II
I


Everything can be integrated by geography
relationships
relationships
temporal
spatial
GIS

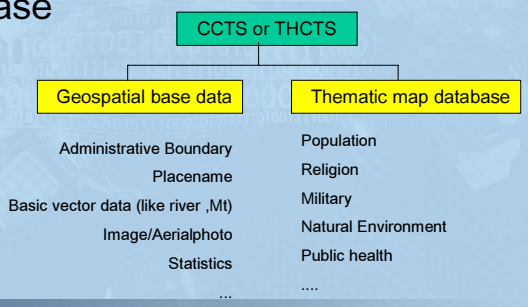
Introduction of the Project

- The CCTS (*Chinese Civilization in Time and Space*) Project which began in 1996 aims to build up a fundamental system for historical GIS and basic data of Chinese history of the past 2000 years. The spatial resolution is about 1:1M~10M.
- The THCTS (*Taiwan History and Culture in Time and Space*) Project which began in 2001 aims to develop a spatial-temporal application infrastructure on the basis of digitalized Taiwanese history, culture and natural resource maps of the past 400 years. The spatial resolution is about 1:20K~100K.

Content of the System

This system consists of three components：

- Geospatial base data
- WebGIS-based application
- Thematic map database



Content of the CCTS System

Chinese Civilization in Time and Space

Geospatial Information

- Admin. Boundaries from HAC
- Inhabited Localities from HAC
- Digital Map Databases of China(ArcChina)
- Digital Chart of the World (DCW)
- Changes of Yellow River
- Pre-historical Cultural Areas
- Gazetteers
- Historical Atlas and Remote Sensing Imagery

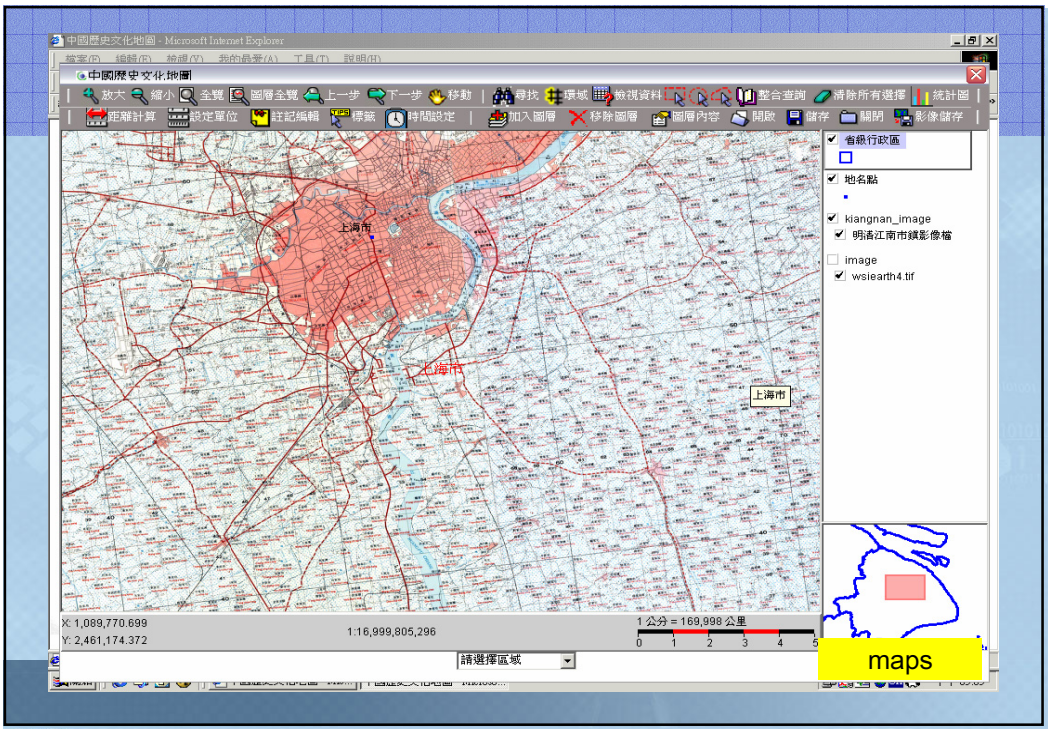
Thematic Information

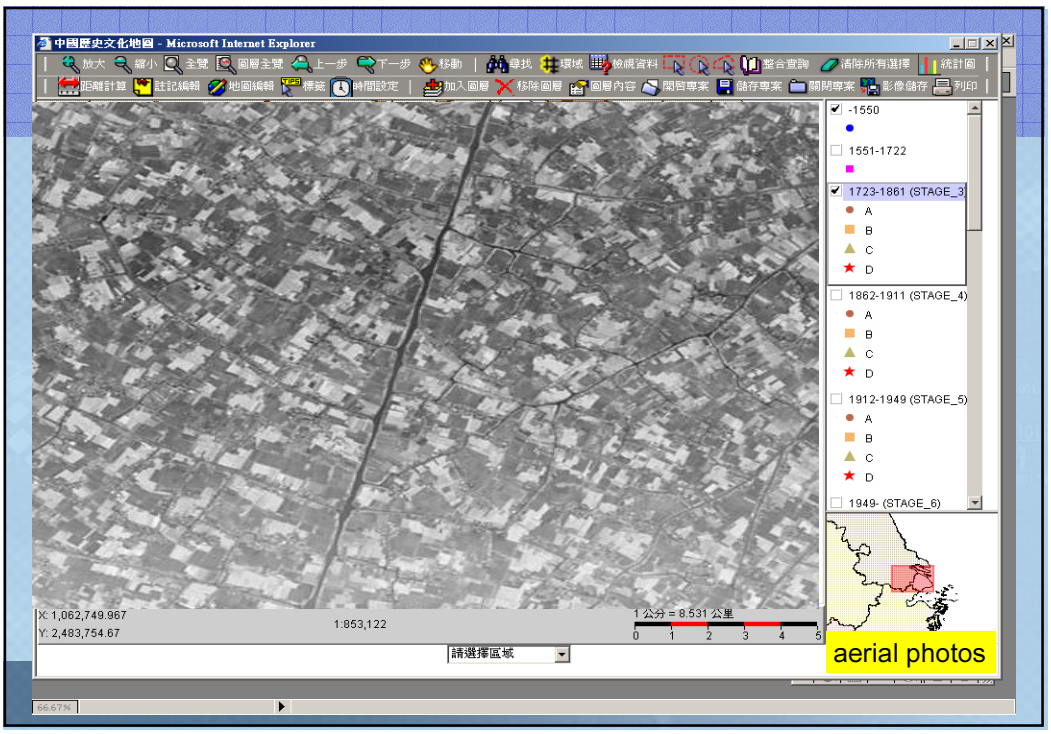
- Scripta Sinica
- Grain Price and Population in Qing
- Beacon Tower Dist. In Eastern Han
- Union Catalog of Local Gazetteers
- DB of Person Names in Chinese History
- Dist. Archaeological Tombs in Han
- DB of Archaeological Sites of Silk Road in Hsinjiang

The Basic Dataset of the CCTS: The Historical Atlas of China

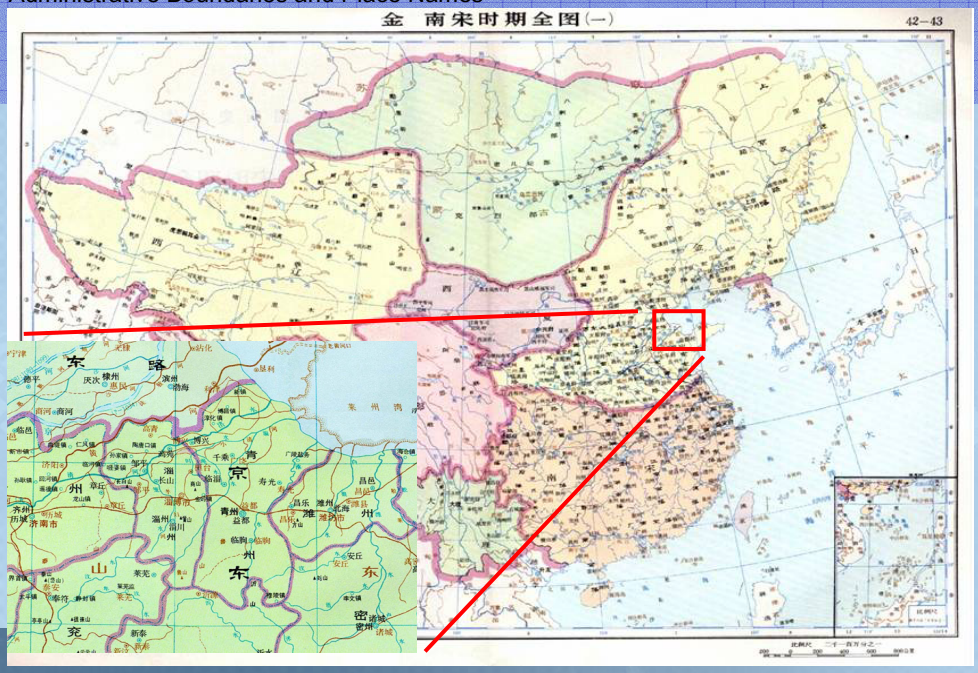
(Various Scales from 1:2,100,000 to 1:7,180,000)

8 Volumes with 20 groups, 304 pieces of Historical Maps from pre-Qin era to the Qing Dynasty

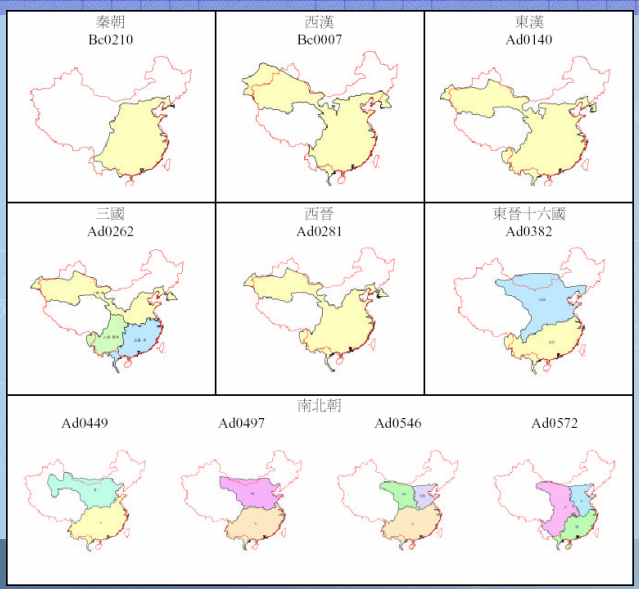

maps


aerial photos

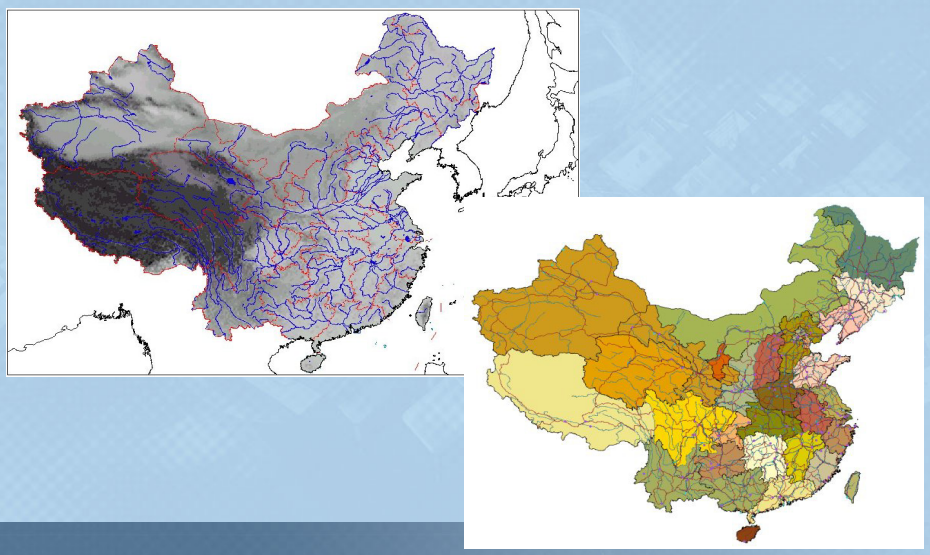
Administrative Boundaries and Place Names



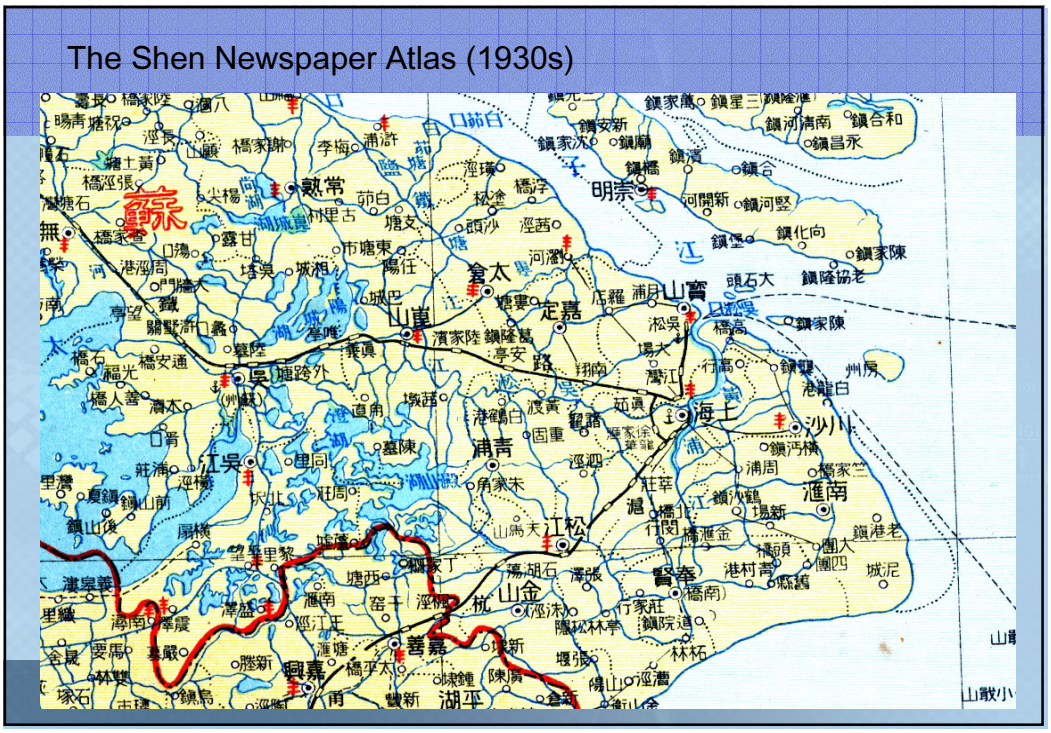
- Changes in Administrative Boundaries

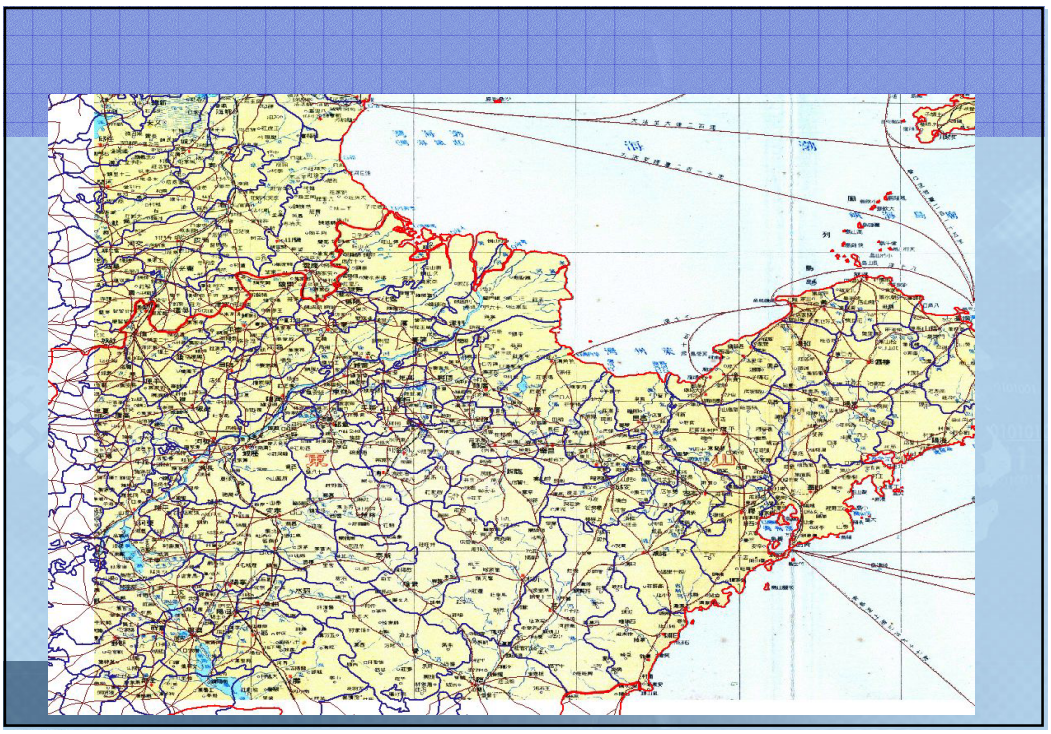


The Basic Dataset of the CCTS: The Digital Map Database of China (Arc/China)(1:1,000,000)



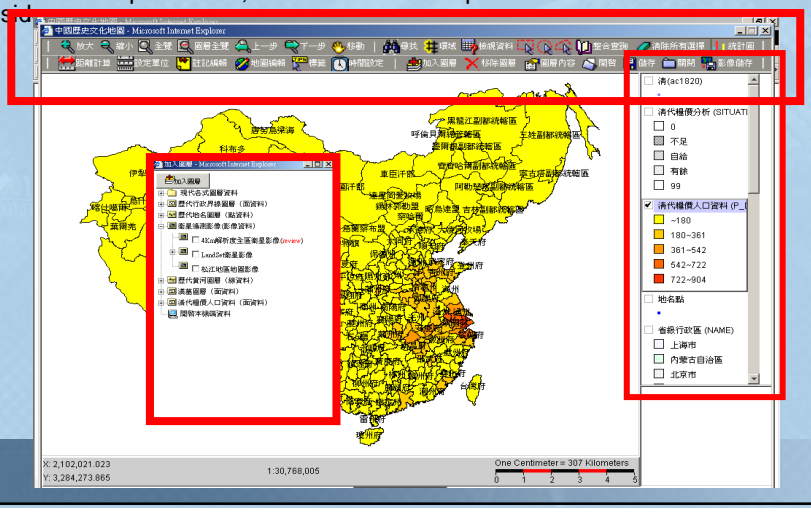
- The Dataset of ArcChina, published by the National Bureau of Surveying and Mapping , PRC., ranges from 1980 to 1990 with 13 features and 77 map blocks,

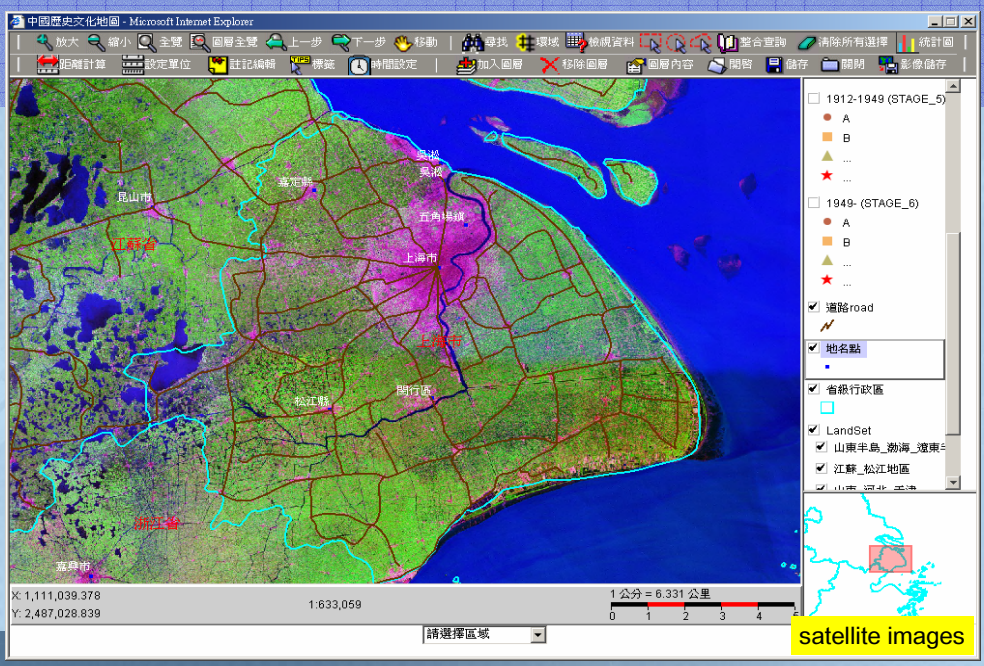

The Shen Newspaper Atlas (1930s)

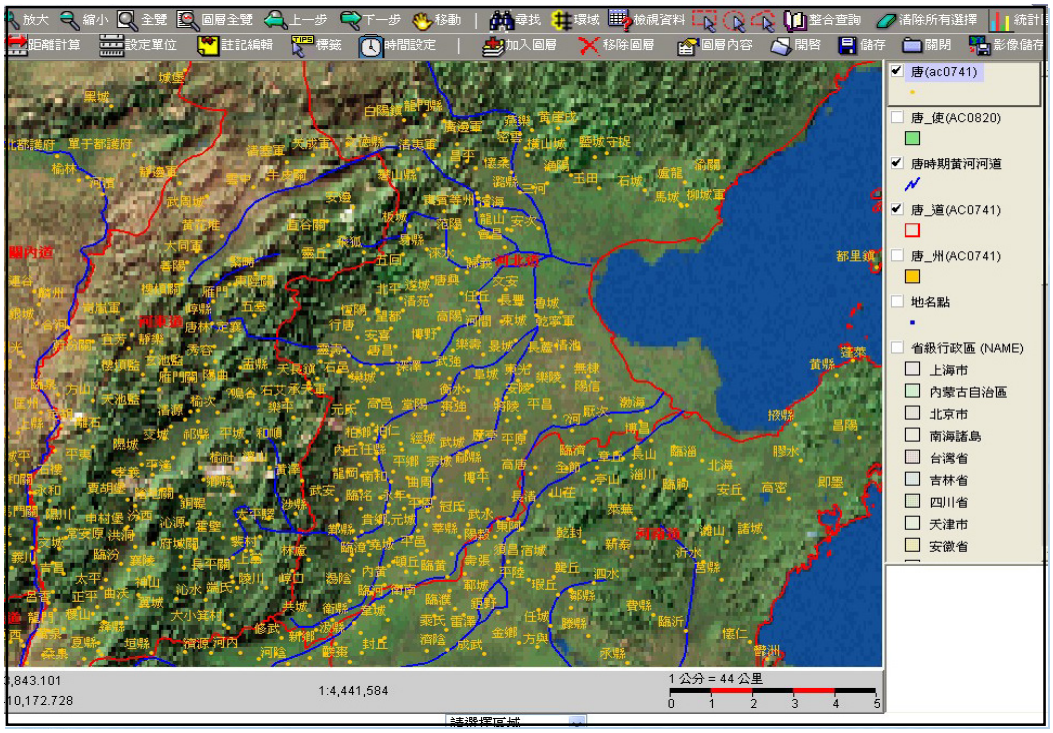


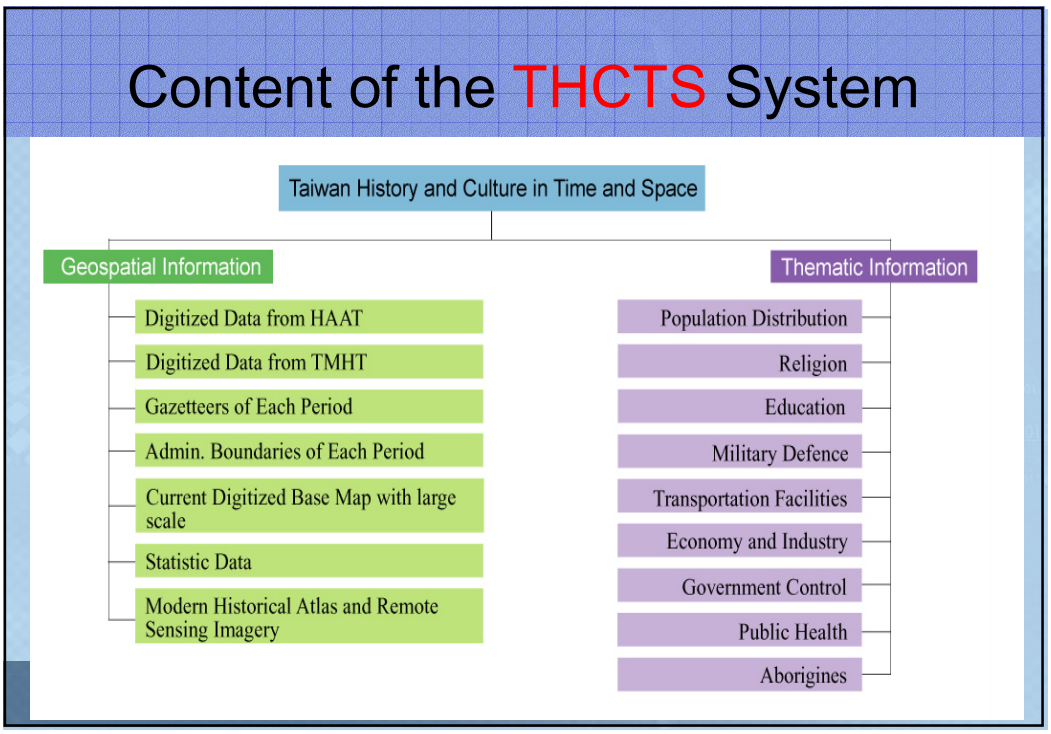
The Chinese Civilization in Time and Space System

1. A WebGIS-based infrastructure for content navigation and web mapping
2. Integrating geospatial and attribute information via the Internet
3. Users can upload data, create their own maps and save them in the client or server side






唐(ac0741)
唐_使(AC0820)
唐時期黃河河道
唐_道(AC0741)
唐_州(AC0741)
地名點
省級行政區 (NAME)
上海市
內蒙古自治區
北京市
南海諸島
台灣省
吉林省
四川省
天津市
安徽省
1:4,441,584
1 公分 = 44 公里

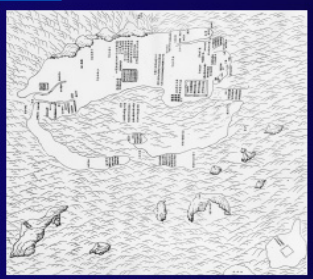

Content of the THCTS System
Taiwan History and Culture in Time and Space
Geospatial Information
Digitized Data from HAAT
Digitized Data from TMHT
Gazetteers of Each Period
Admin. Boundaries of Each Period
Current Digitized Base Map with large scale
Statistic Data
Modern Historical Atlas and Remote Sensing Imagery
Thematic Information
Population Distribution
Religion
Education
Military Defence
Transportation Facilities
Economy and Industry
Government Control
Public Health
Aborigines

◦ Geospatial base maps of THCTS

◦ Taiwanese history can be divided into: Dutch and Spanish Period, Chengsí Dynasty, Qing Dynasty, Japanese Colonized Period and After WW II. Accordingly, we have developed various base maps of each period. Among them,《Atlas of the historic administrative pau division of Taiwan》(1904) ,《Atlas of the topography of Taiwan during the period of Japanese rule》(1920) and the newly published, 《 Topography of Taiwan 》, photographic maps and remote sensing images, constitute the main body of base maps.

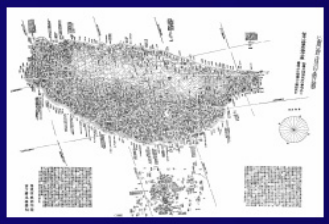
17th~19th old maps

康熙五十三年(1714) 測繪之臺灣地圖



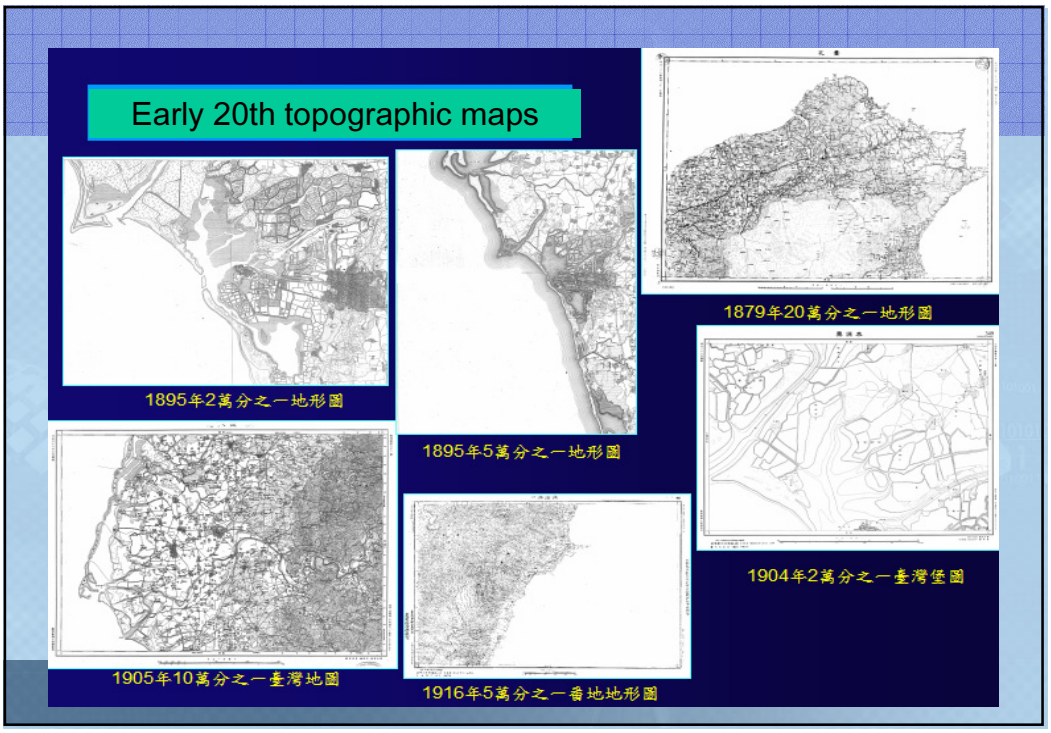
永曆十八年(1664)臺灣軍備圖

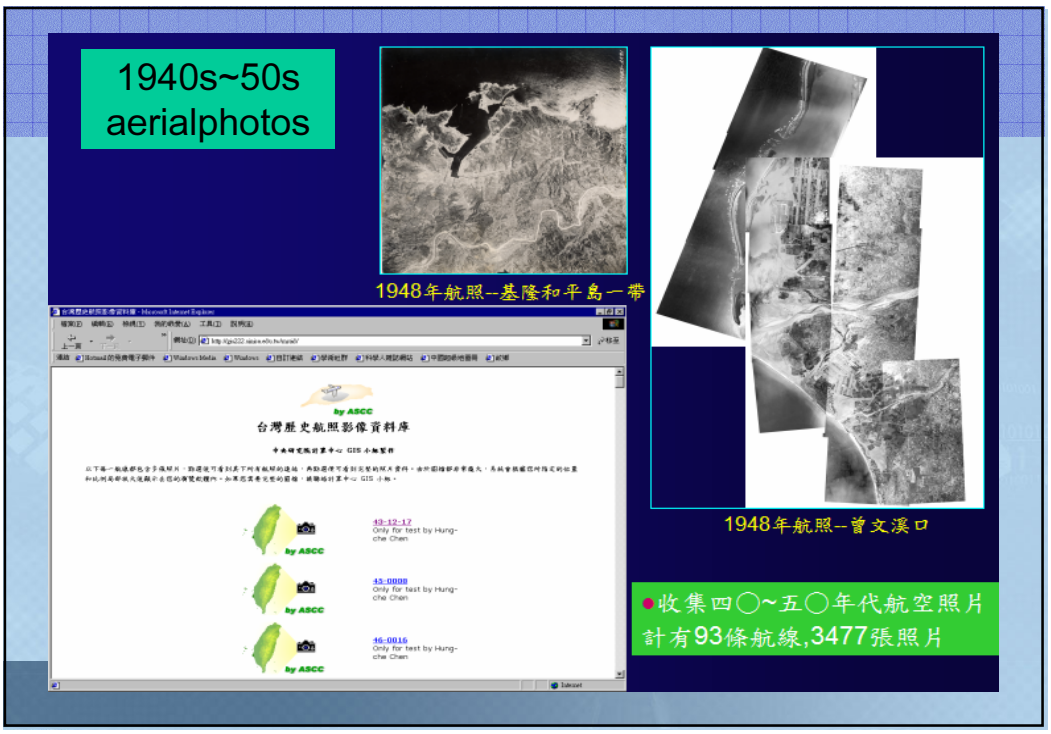
廣東福建與臺灣海圖[舉南圖](1650-1660)

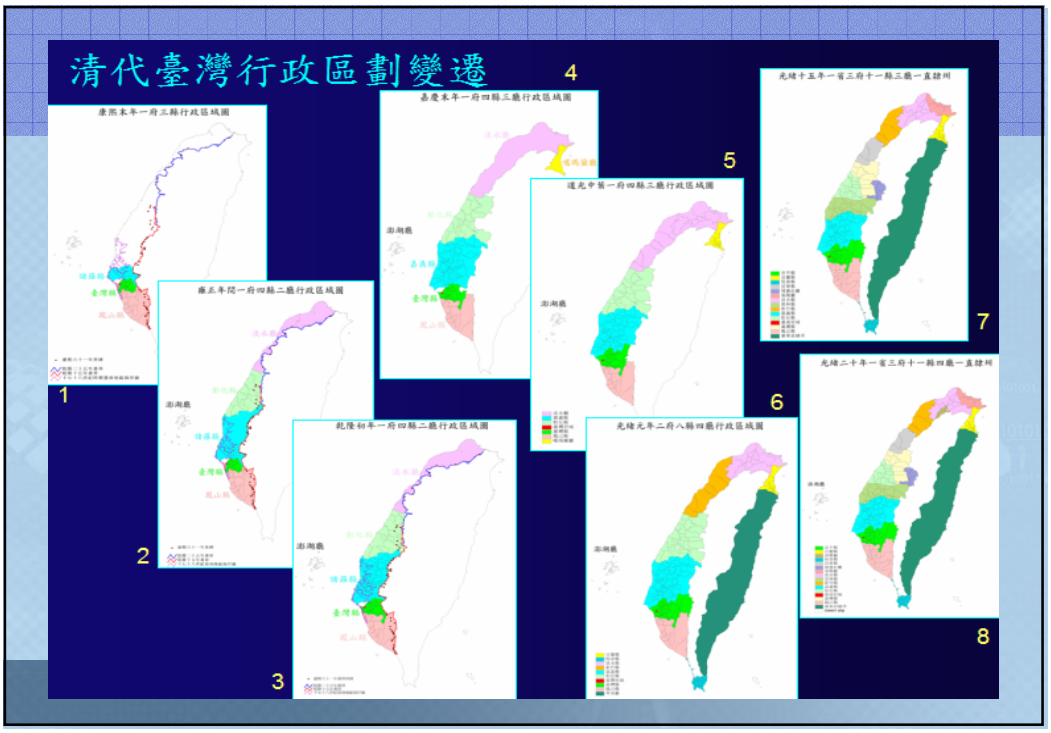


光緒五年(1879)臺灣後山總圖

清康熙中葉臺灣輿圖

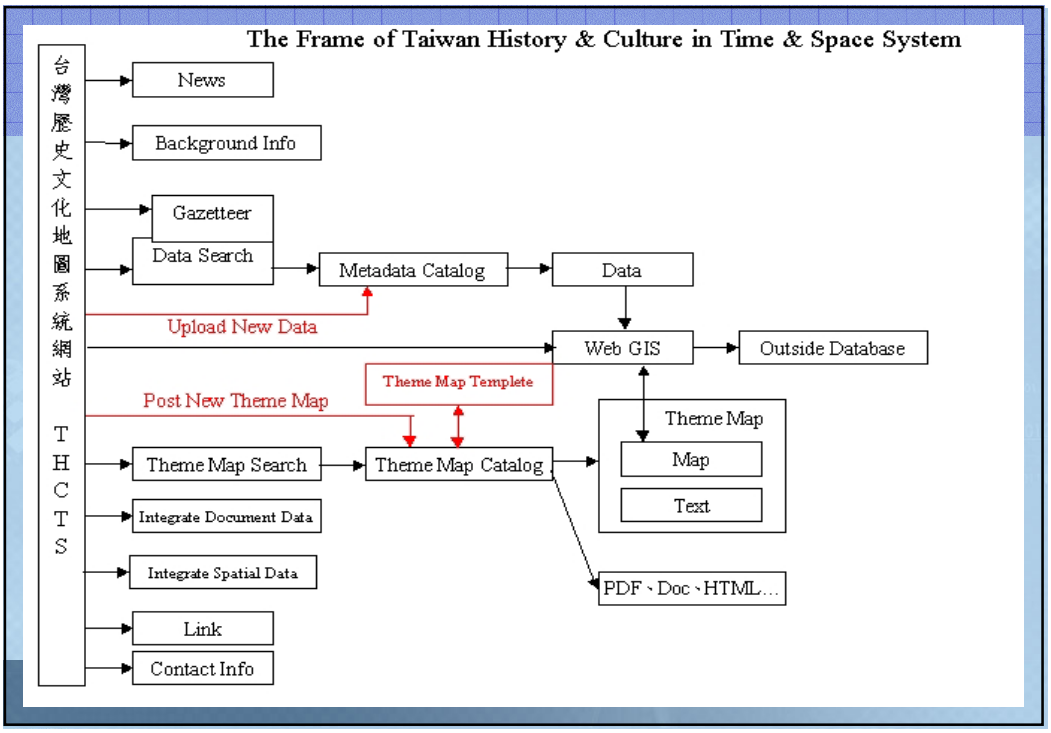

Early 20th topographic maps
1895年2萬分之一地形圖
1895年5萬分之一地形圖
1879年20萬分之一地形圖
1904年2萬分之一臺灣堡圖
1905年10萬分之一臺灣地圖
1916年5萬分之一番地地形圖

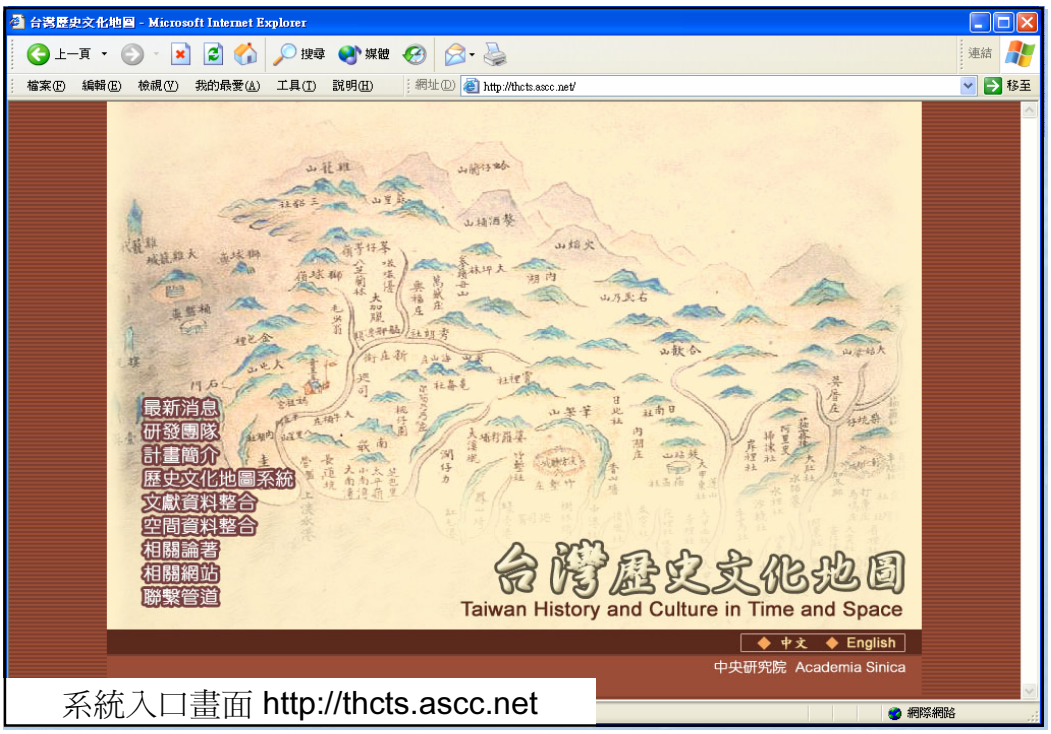

1940s~50s aerialphotos
1948年航照--基隆和平島一帶
1948年航照--曾文溪口
台灣歷史航照影像資料庫
●收集四〇~五〇年代航空照片
計有93條航線,3477張照片



◦ Thematic Map Database of THCTS

We have integrated available researches in Taiwanese history and culture with base maps and system application infrastructure. Nine groups of thematic maps were produced, such as population distribution, religion, education and aboriginal people. In the future, we hope to make more value of this spatial-temporal infrastructure through active participation from scholars across different fields.



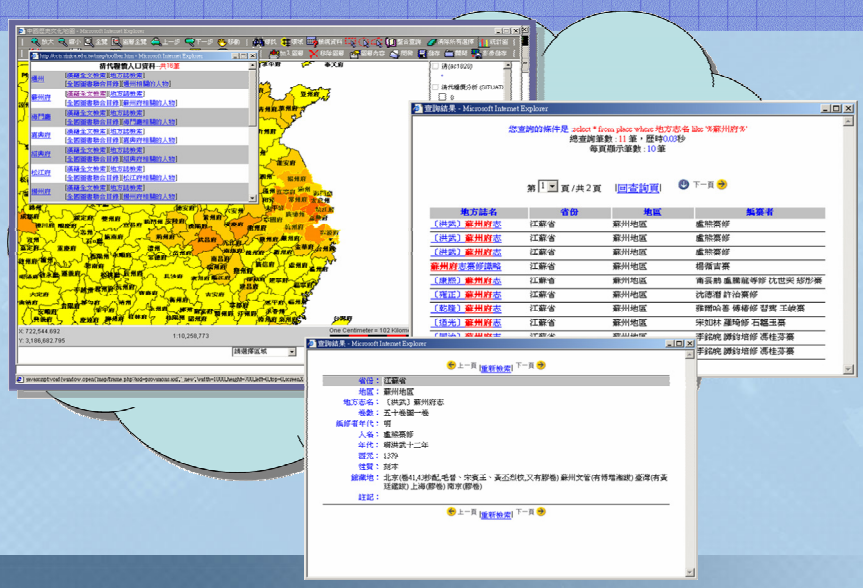


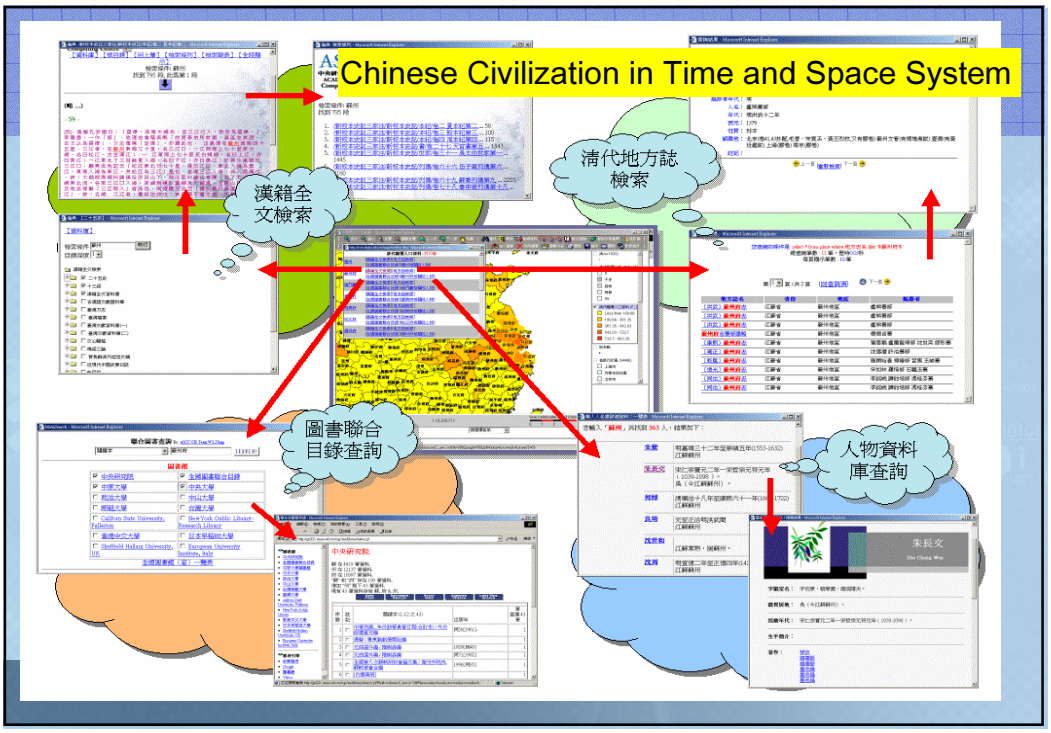
系統入口畫面 http://thcts.ascc.net

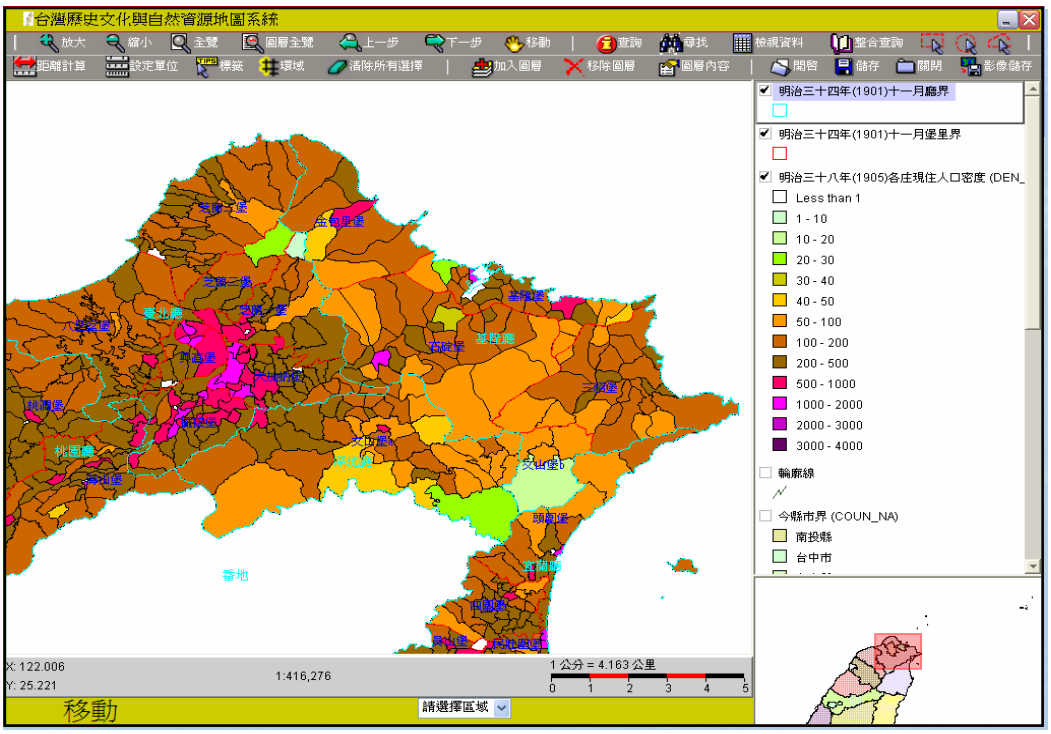
Applications of CCTS

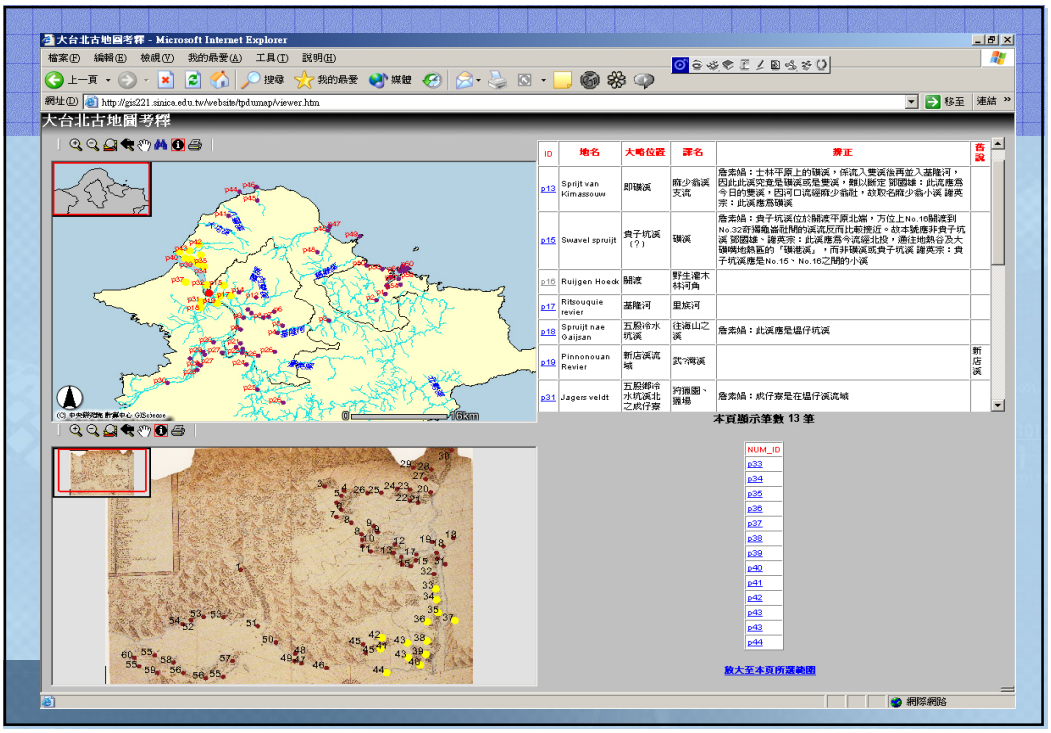
Integration and Retrieval of Data

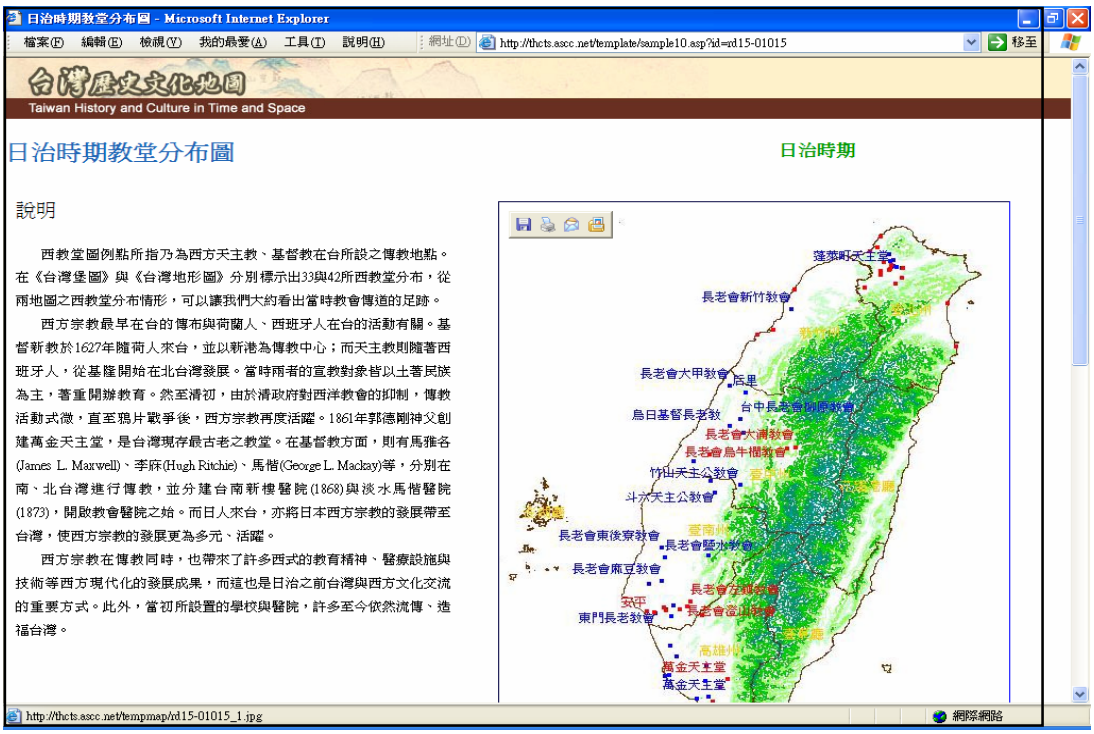
Searching the Bibliography of Chinese Local Gazetteers (Local History)

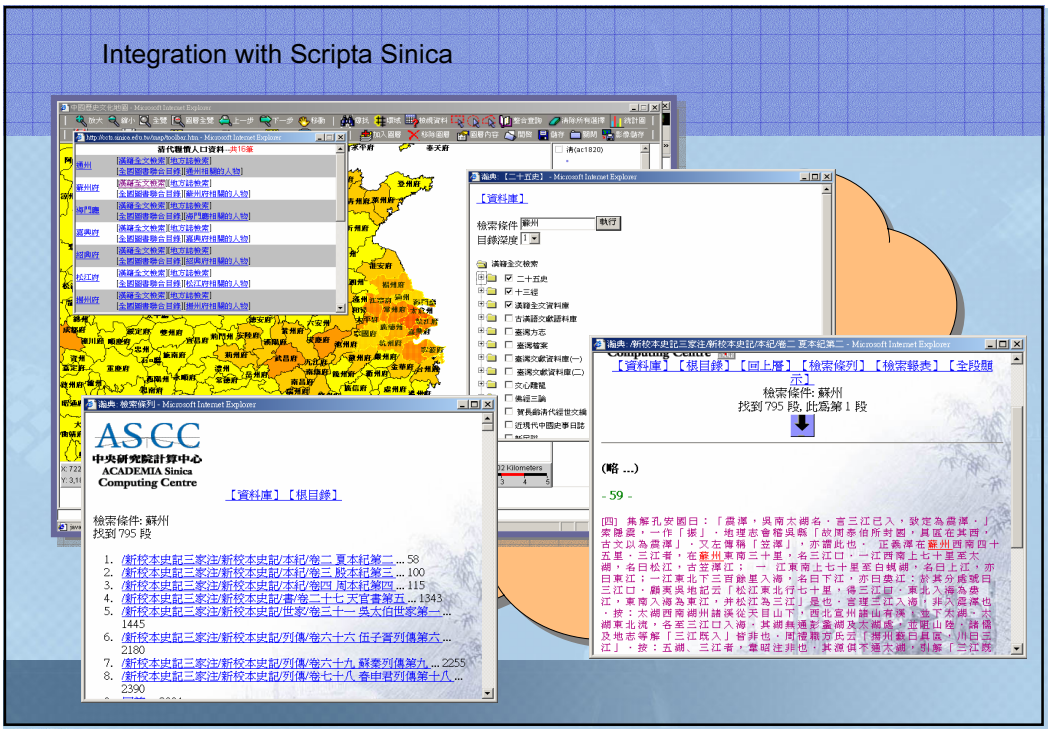


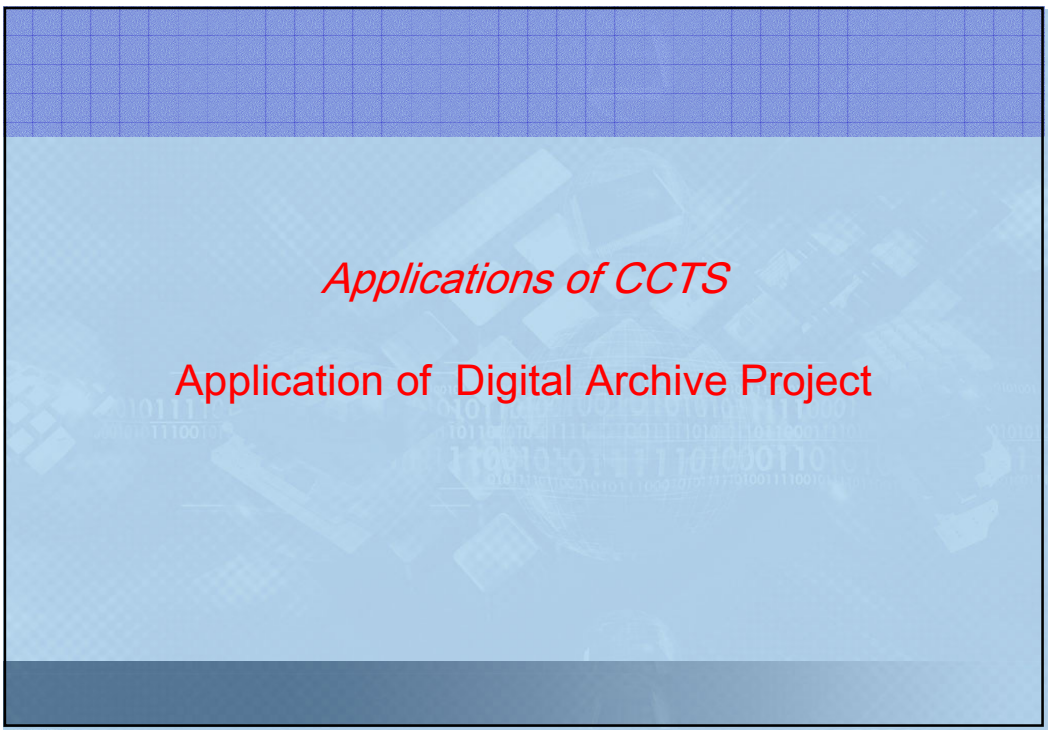

Chinese Civilization in Time and Space System
漢籍全文檢索
清代地方誌檢索
圖書聯合目錄查詢
人物資料庫查詢

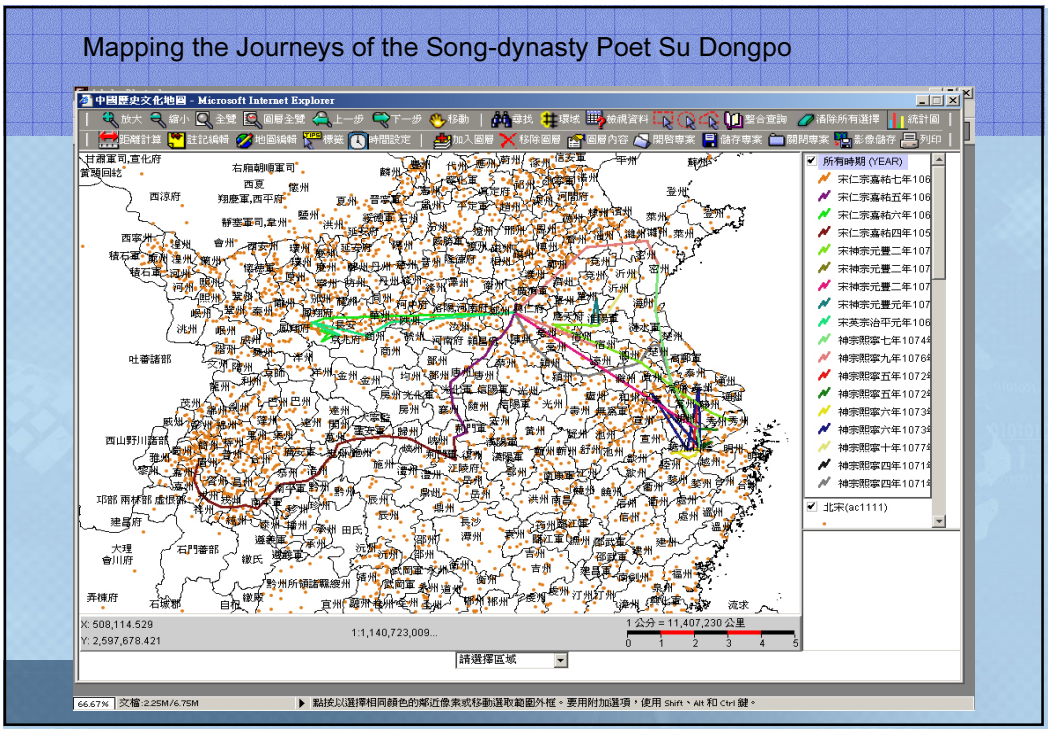

台灣歷史文化與自然資源地圖系統
放大
縮小
全覽
圖層全覽
上一步
下一步
移動
查詢
尋找
檢視資料
整合查詢
距離計算
設定單位
標籤
環域
清除所有選擇
加入圖層
移除圖層
圖層內容
開啓
儲存
關閉
影像儲存
明治三十四年(1901)十一月廳界
明治三十四年(1901)十一月堡里界
明治三十八年(1905)各庄現住人口密度 (DEN_
Less than 1
1 - 10
10 - 20
20 - 30
30 - 40
40 - 50
50 - 100
100 - 200
200 - 500
500 - 1000
1000 - 2000
2000 - 3000
3000 - 4000
輪廓線
今縣市界 (COUN_NA)
南投縣
台中市
X: 122.006
Y: 25.221
1:416,276
1 公分 = 4.163 公里
移動
請選擇區域

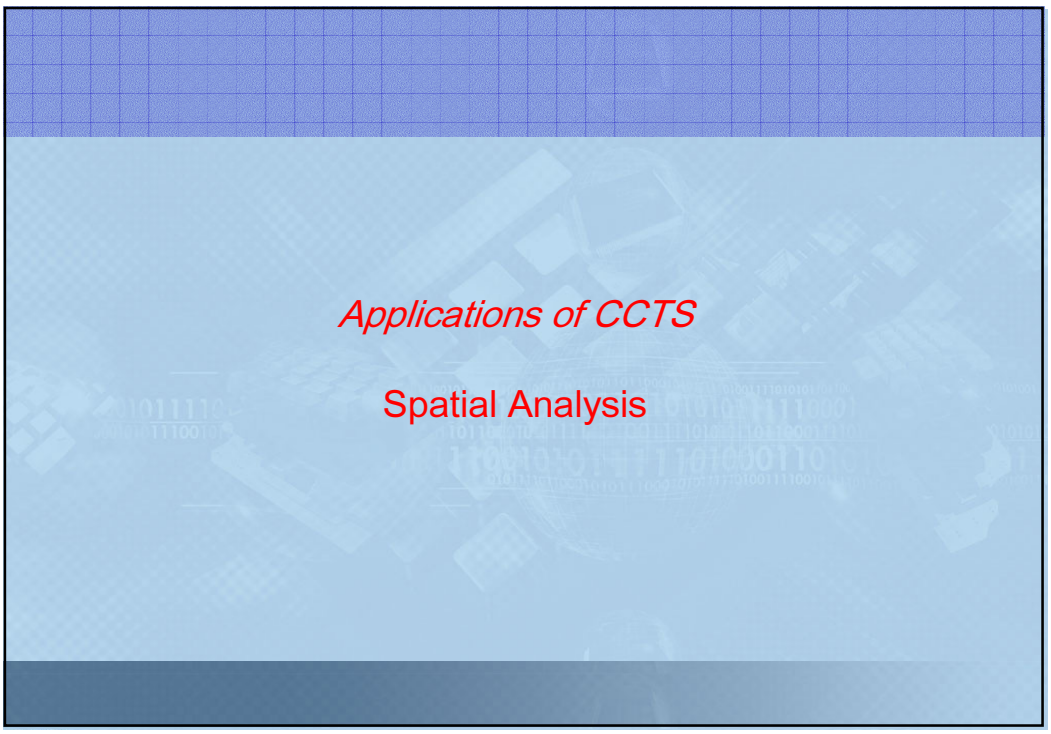

大台北古地圖考釋 - Microsoft Internet Explorer
大台北古地圖考釋
本頁顯示筆數 13 筆

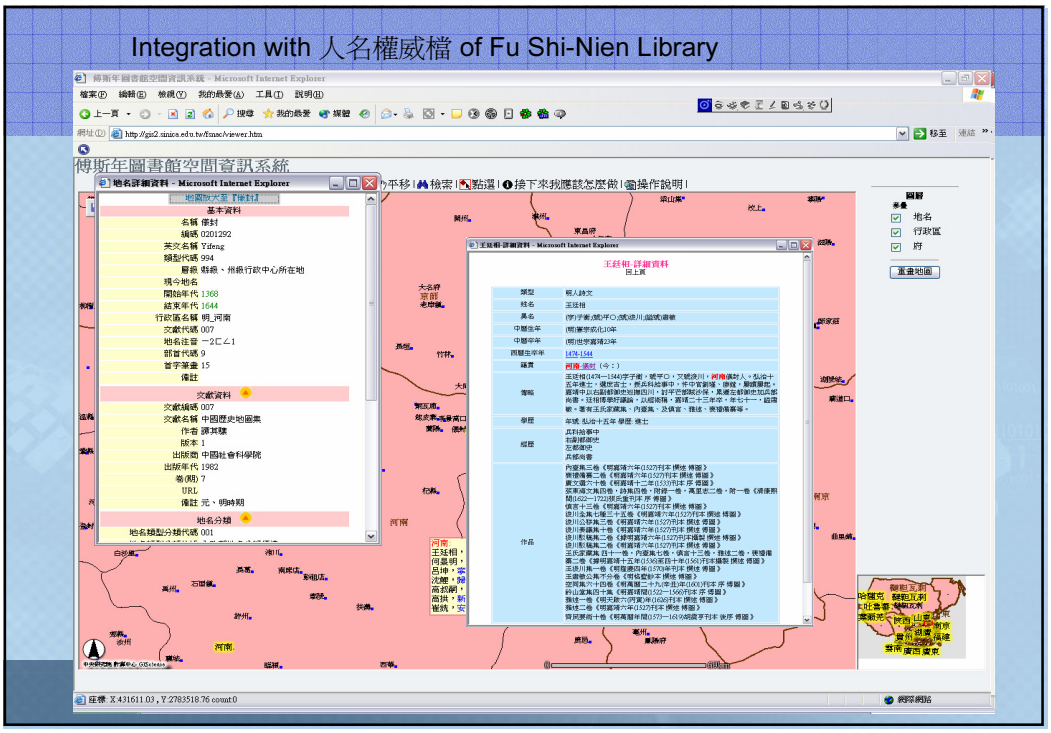

日治時期教堂分布圖 - Microsoft Internet Explorer
台灣歷史文化地圖
Taiwan History and Culture in Time and Space
日治時期教堂分布圖
日治時期
說明
西教堂圖例點所指乃為西方天主教、基督教在台所設之傳教地點。在《台灣堡圖》與《台灣地形圖》分別標示出33與42所西教堂分布，從兩地圖之西教堂分布情形，可以讓我們大約看出當時教會傳道的足跡。
西方宗教最早在台的傳布與荷蘭人、西班牙人在台的活動有關。基督新教於1627年隨荷人來台，並以新港為傳教中心；而天主教則隨著西班牙人，從基隆開始在北台灣發展。當時兩者的宣教對象皆以土著民族為主，著重開辦教育。然至清初，由於清政府對西洋教會的抑制，傳教活動式微，直至鴉片戰爭後，西方宗教再度活躍。1861年郭德剛神父創建萬金天主堂，是台灣現存最古老之教堂。在基督教方面，則有馬雅各(James L. Maxwell)、李庥(Hugh Ritchie)、馬偕(George L. Mackay)等，分別在南、北台灣進行傳教，並分建台南新樓醫院(1868)與淡水馬偕醫院(1873)，開啟教會醫院之始。而日人來台，亦將日本西方宗教的發展帶至台灣，使西方宗教的發展更為多元、活躍。
西方宗教在傳教同時，也帶來了許多西式的教育精神、醫療設施與技術等西方現代化的發展成果，而這也是日治之前台灣與西方文化交流的重要方式。此外，當初所設置的學校與醫院，許多至今依然流傳、造福台灣。
蓬萊町天主堂
長老會新竹教會
長老會大甲教會
烏日基督長老教
長老會大湳教會
竹山天主公教會
斗六天主公教會
長老會東後寮教會
長老會麻豆教會
東門長老教會
萬金天主堂
萬金天主堂

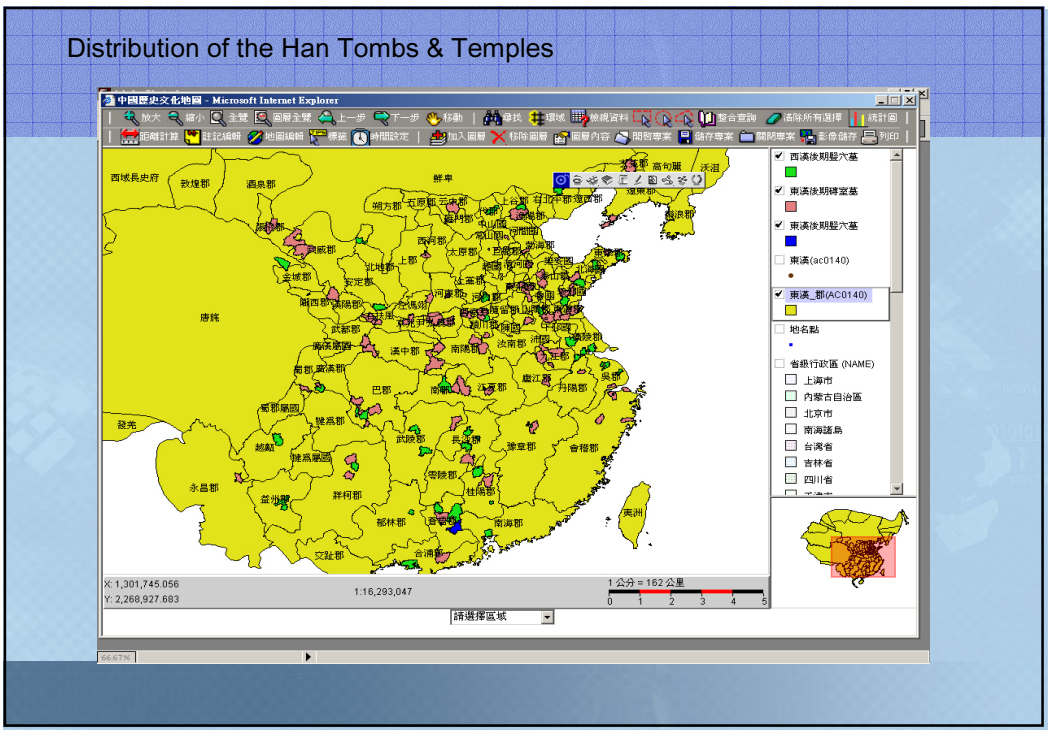

Integration with Scripta Sinica
ASCC
中央研究院計算中心
ACADEMIA Sinica
Computing Centre
【資料庫】【根目錄】
檢索條件: 蘇州
找到 795 段
1. /新校本史記三家注/新校本史記/本紀/卷二 夏本紀第二... 58
2. /新校本史記三家注/新校本史記/本紀/卷三 殷本紀第三... 100
3. /新校本史記三家注/新校本史記/本紀/卷四 周本紀第四... 115
4. /新校本史記三家注/新校本史記/書/卷二十七 天官書第五... 1343
5. /新校本史記三家注/新校本史記/世家/卷三十一 吳太伯世家第一... 1445
6. /新校本史記三家注/新校本史記/列傳/卷六十六 伍子胥列傳第六... 2180
7. /新校本史記三家注/新校本史記/列傳/卷六十九 蘇秦列傳第九... 2255
8. /新校本史記三家注/新校本史記/列傳/卷七十八 春申君列傳第十八... 2390
【資料庫】
檢索條件: 蘇州
目錄深度
【資料庫】【根目錄】【回上層】【檢索條列】【檢索報表】【全段顯示】
檢索條件: 蘇州
找到 795 段, 此爲第 1 段
(略 ...)
- 59 -


Applications of CCTS
Application of Digital Archive Project

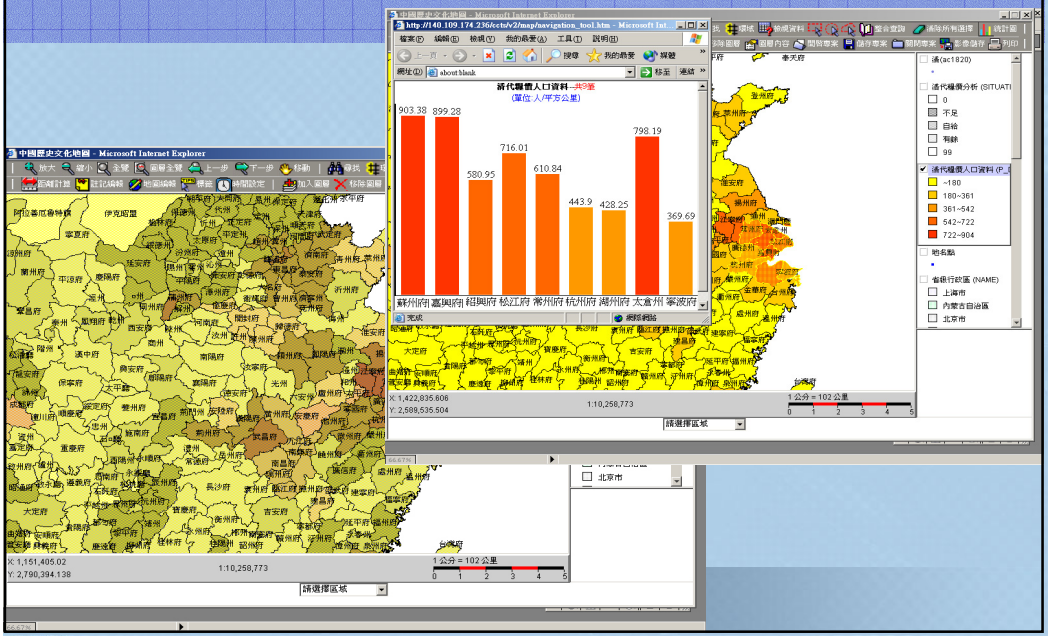

Mapping the Journeys of the Song-dynasty Poet Su Dongpo
中國歷史文化地圖 - Microsoft Internet Explorer


Applications of CCTS
Spatial Analysis

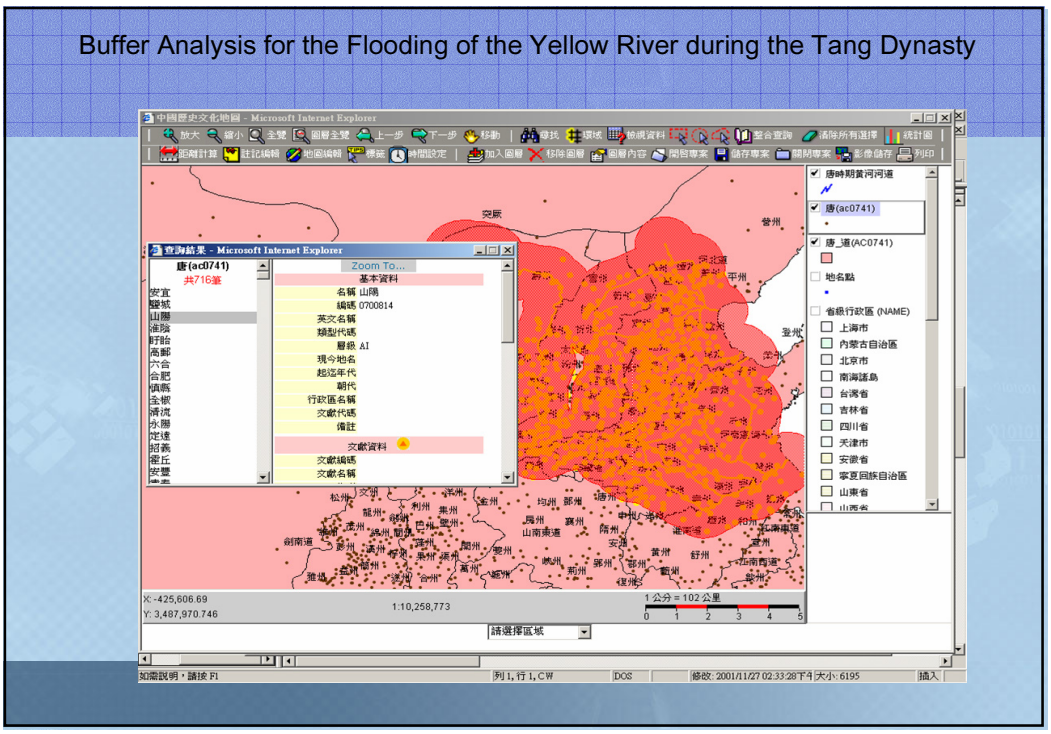

Integration with 人名權威檔 of Fu Shi-Nien Library
傅斯年圖書館空間資訊系統


Distribution of the Han Tombs & Temples
中國歷史文化地圖 - Microsoft Internet Explorer

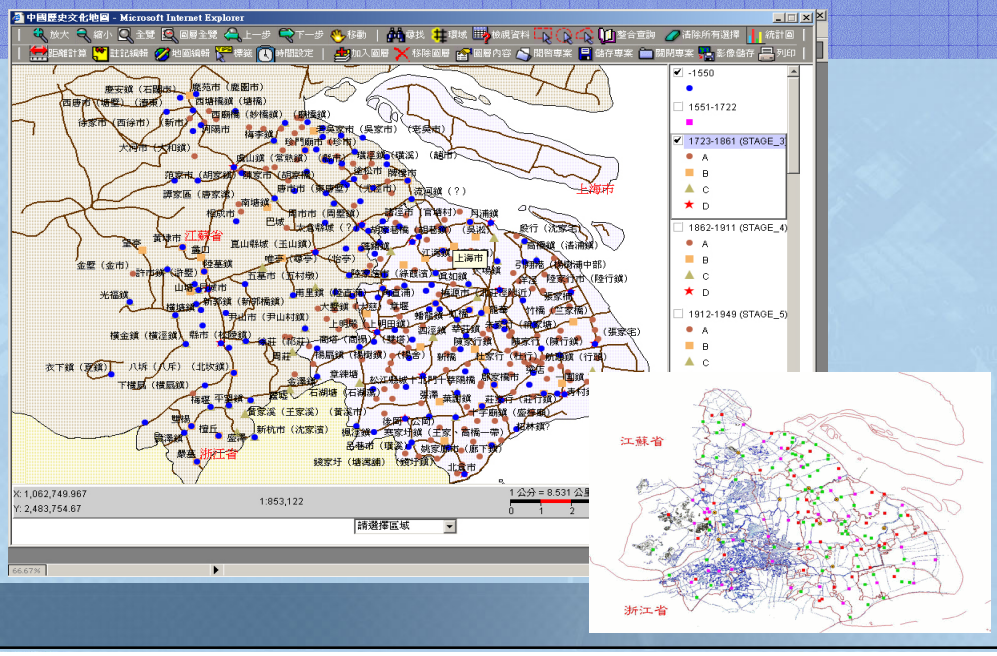
Price Trends and Population Distribution during the Qing Dynasty



Buffer Analysis for the Flooding of the Yellow River during the Tang Dynasty

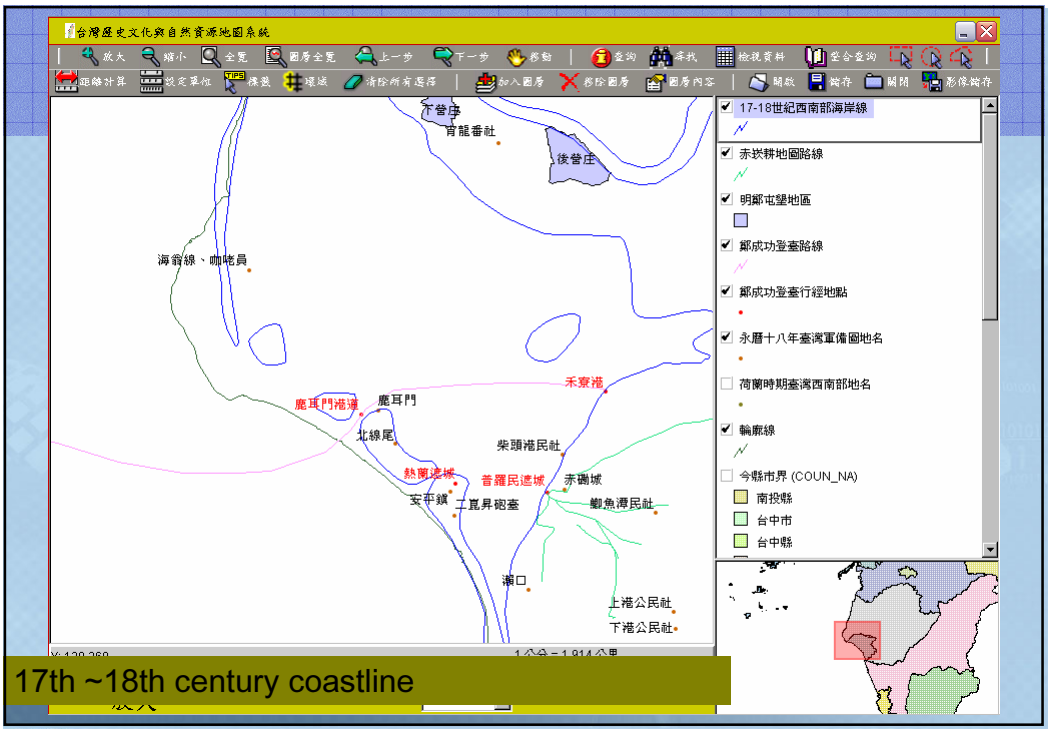


Overlay Analysis of the Ming-Qing Jiangnan Market Towns

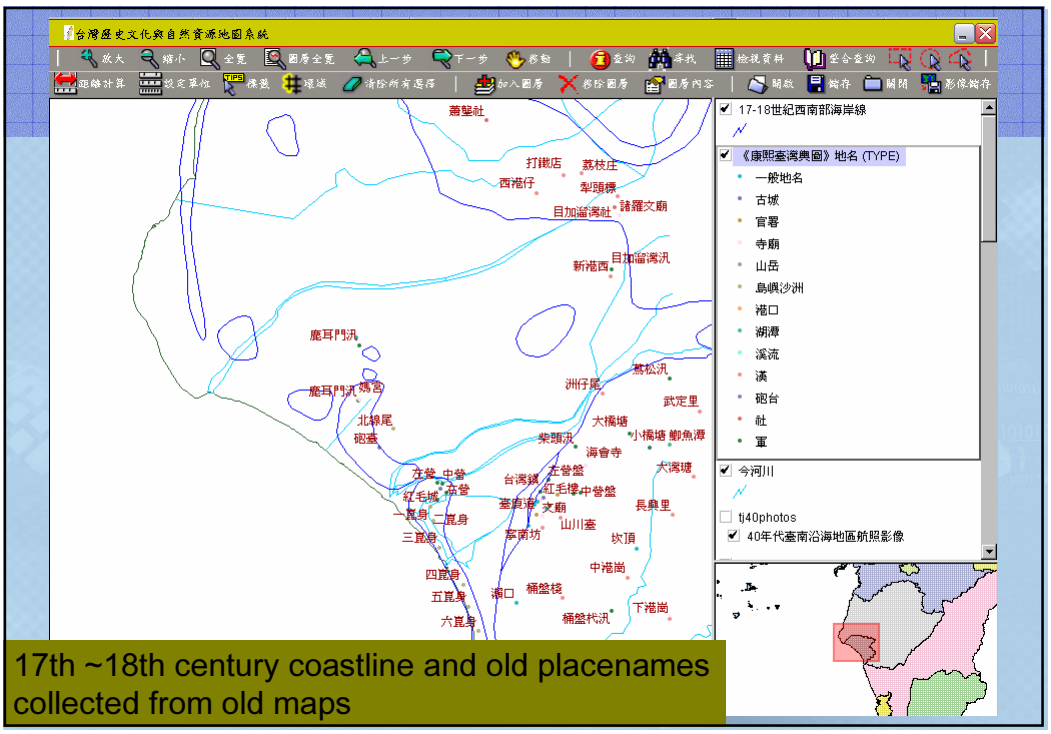


Applications of THCTS

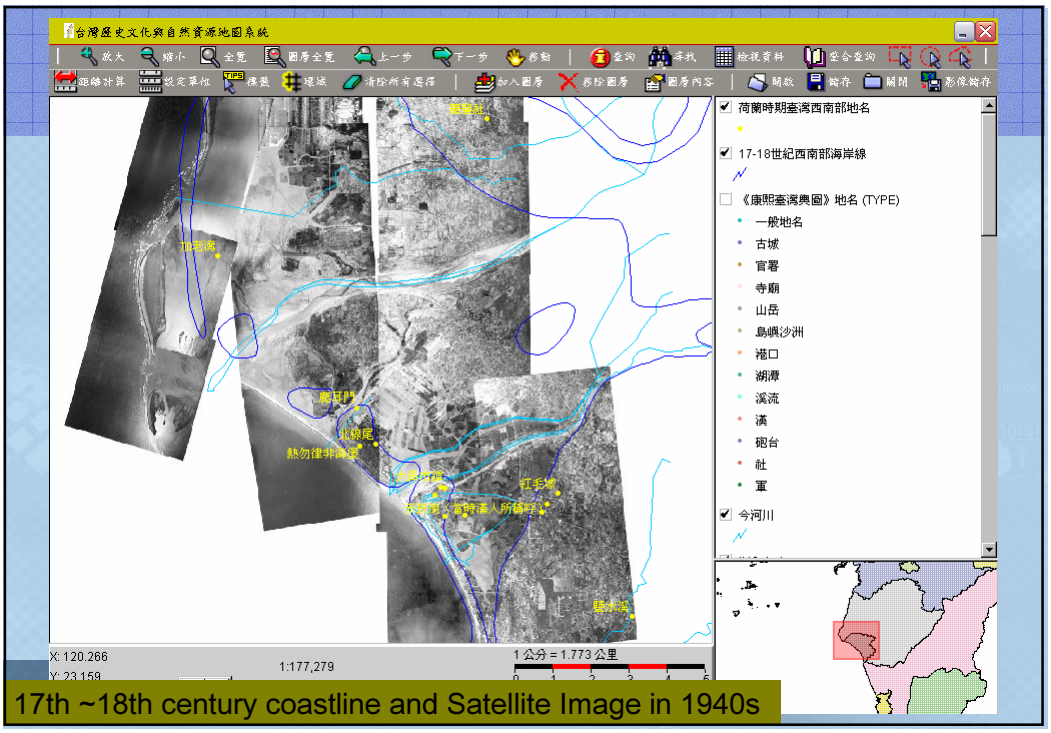
Area Study: Tainan, Taiwan as an example



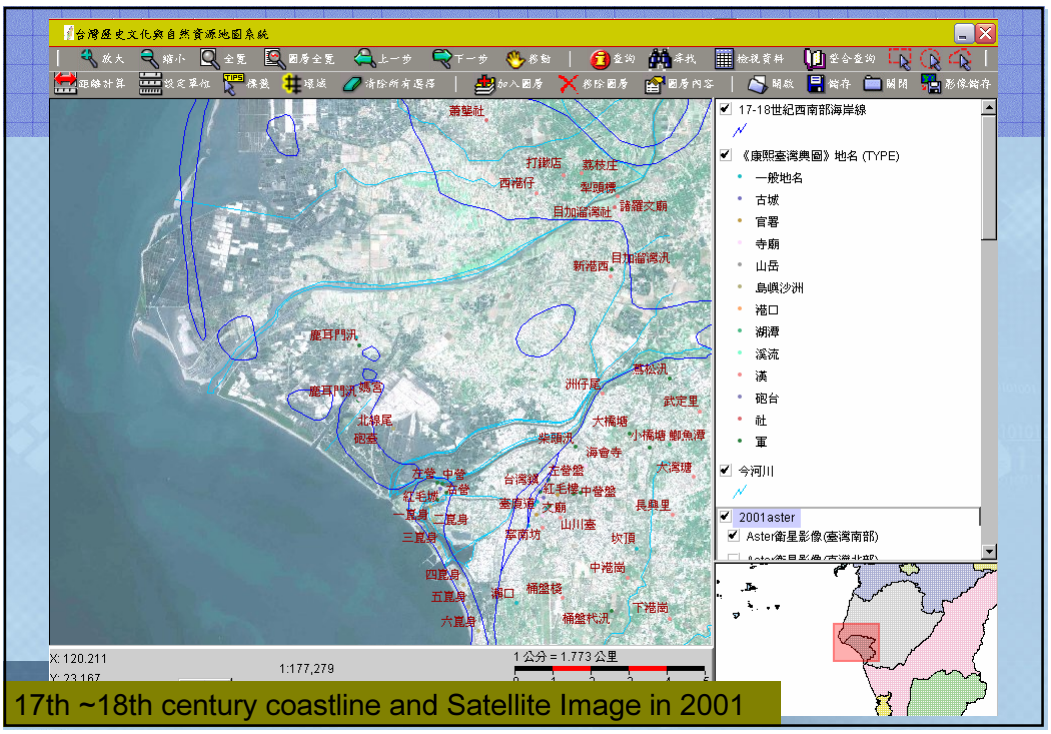
17th ~18th century coastline



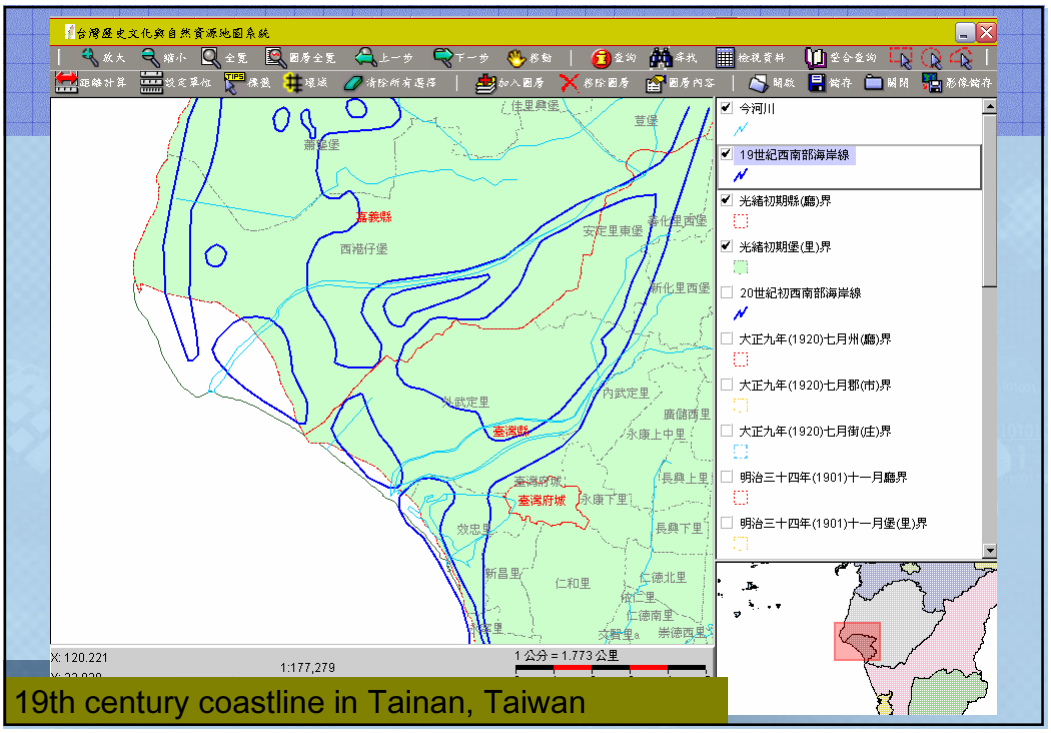
17th ~18th century coastline and old placenames collected from old maps



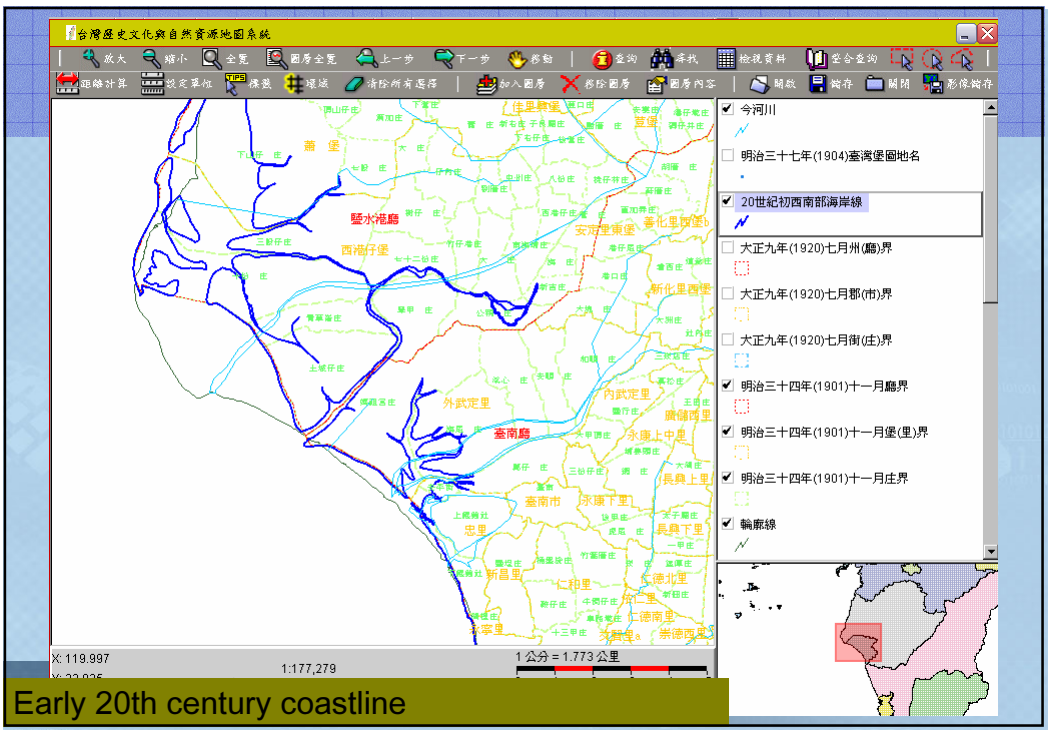
17th ~18th century coastline and Satellite Image in 1940s



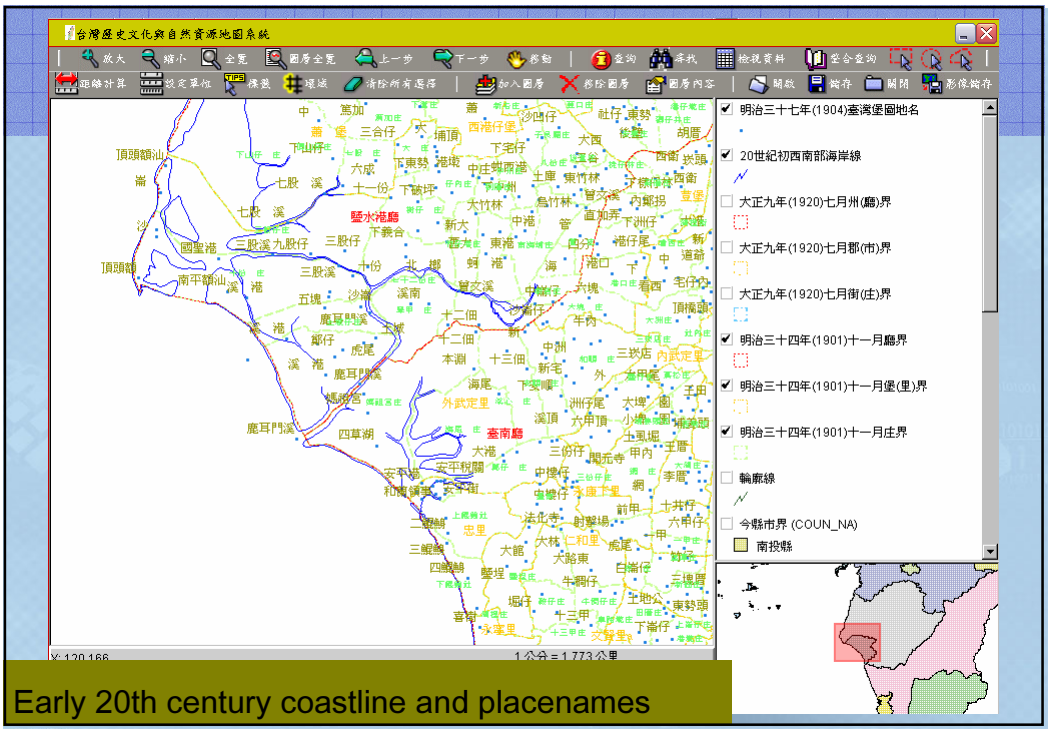
17th ~18th century coastline and Satellite Image in 2001



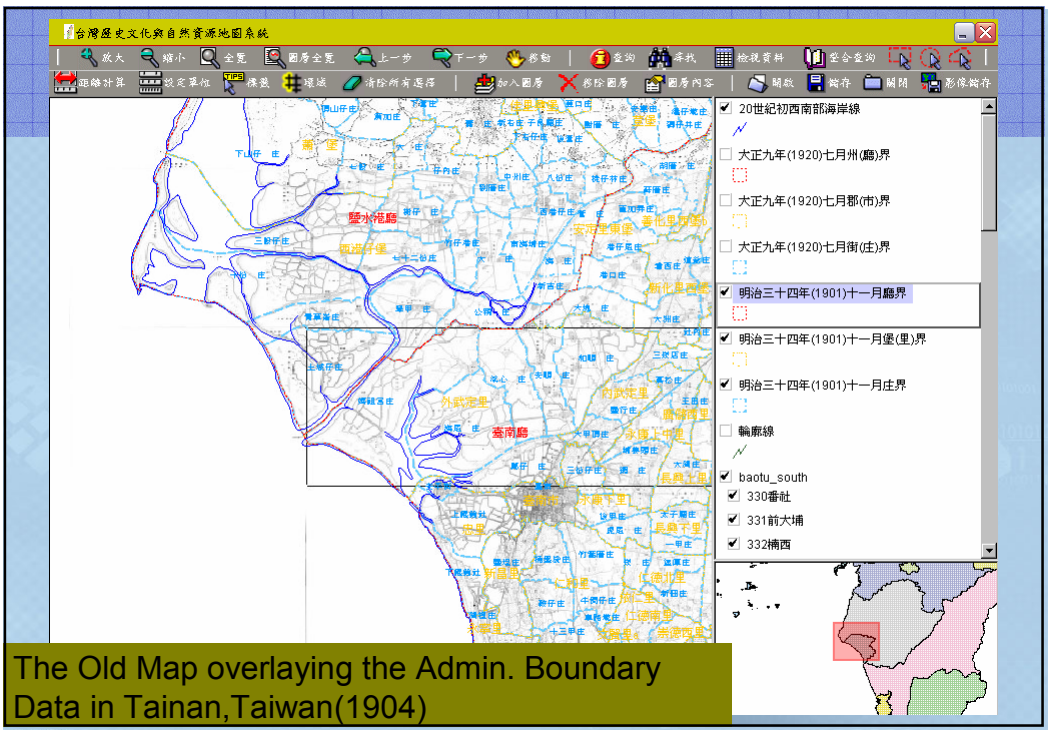
19th century coastline in Tainan, Taiwan



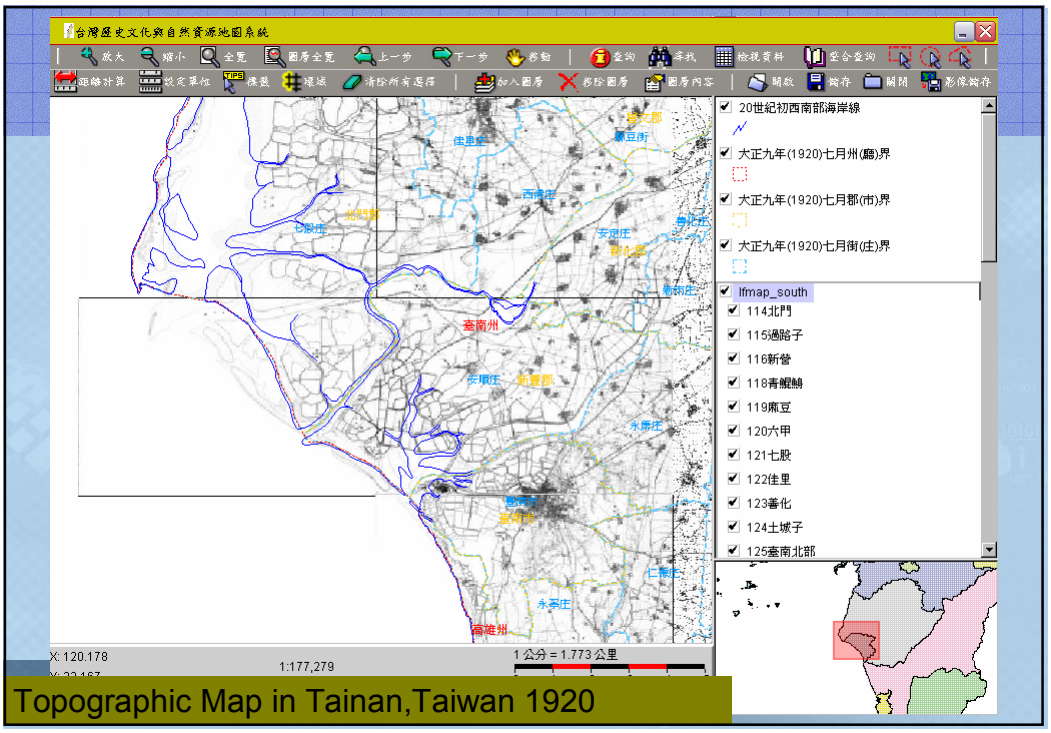
Early 20th century coastline



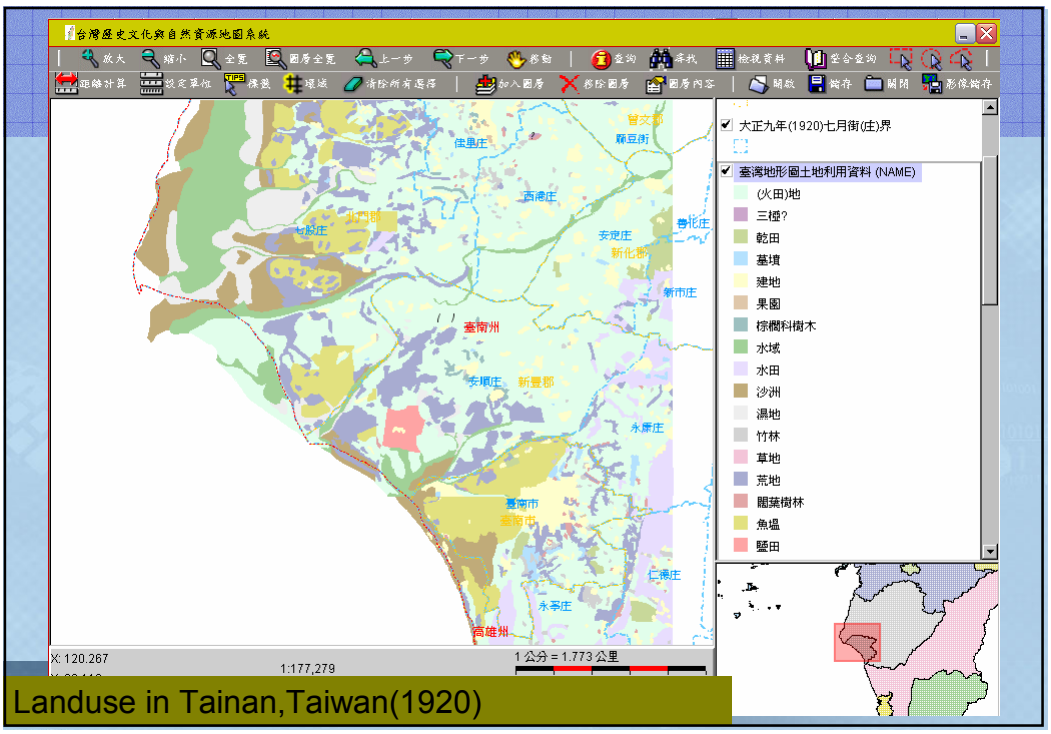
Early 20th century coastline and placenames



The Old Map overlaying the Admin. Boundary Data in Tainan,Taiwan(1904)



Topographic Map in Tainan,Taiwan 1920



Landuse in Tainan,Taiwan(1920)

Summary

- The CCTS (*Chinese Civilization in Time and Space*) Project covers a large spatial and temporal extent and builds up historic features using available historic geography study.
- The THCTS (*Taiwan History and Culture in Time and Space*) Project has high spatial and temporal resolution and builds up historic features from first-hand historic studies using GIS. The thematic maps are distinguishing features.

Current Work

- *Collecting more data*
- *Developing integration Interface to ECAI cleanhouse and other Web-GIS portal sites with standard protocol*
- *Promoting the utilization of both systems*
  - *UC Berkley and Colombia Uni. have installed CCTS.*
  - *Major universities in Taiwan have been authorized access to both systems.*
- *Working with National Digital Archives Project in Taiwan*
- *Developing Web-GIS client-side display environment using GML and SVG*
- *Testing Web-Editing Function in CCTS and THCTS*
- *Porting Web-GIS to Grid services*

Future Prospect

- *Open Lab for China and Taiwan Studies*
  - Creation of customized GIS project
- *Geolibrary*
  - Cross-cultural and multi-lingual compatibility of formats
  - Integration of information system
  - Clearinghouse & Protocol for the domain knowledge of China and Taiwan studies
- *Virtual Center of Sinological Studies*

Contact Information

- CCTS URL: http://ccts.ascc.net
- THCTS URL: http://thcts.ascc.net
- Other Projects: http://gis.ascc.net/
  - ccts@sinica.edu.tw
  - I-Chun Fan
    - mhfanbbc@ccvax.sinica.edu.tw
  - Hsuing-Ming Liao
    - veevee@gate.sinica.edu.tw

Sinica BOW and 300 Tang Poems:

An overview of a bilingual ontological wordnet and its application to a small ontology of Tang poetry

研究院知識詞網與唐詩三百首

—雙語知識本體詞網簡介及唐詩知識本體之初步構建

Chu-Ren Huang (Academia Sinica), Feng-ju Lo (Yuan Ze University),
Ru-Yng Chang (Academia Sinica), Sueming Chang (Academia Sinica)
黃居仁（中央研究院），羅鳳珠（元智大學），張如瑩（中央研究院），張舒茗（中央研究院）

Abstract

The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED), and SUMO (Suggested Upper Merged Ontology). The three resources were originally linked in two pairs: WordNet 1.6 was manually mapped to SUMO (Niles and Pease 2003) and also to ECTED (the English lemmas in WordNet were mapped to their Chinese lexical equivalents). ECTED encodes both equivalent pairs and their semantic relations (Huang et al. 2003). With the integration of these three key resources, Sinica BOW functions both as an English-Chinese bilingual wordnet and a bilingual lexical access to SUMO.

Sinica BOW allows versatile access and facilitates a combination of lexical, semantic, and ontological information. Versatility is built in with its bilinguality, and the lemma-based merging of multiple resources. First, either English or Chinese can be used for the query, as well as for presenting the content of the resources. Second, the user can easily access the logical structure of both the WordNet and SUMO ontology using either words or conceptual nodes. That is, users can use words to search for ontology or use ontological nodes to search for linguistic realizations in both languages. Third, multiple linguistic indexing is built in to allow additional versatility. Fourth, domain information allows another dimension of knowledge manipulation.

In addition to serving as the reference and infrastructure for the construction of specific knowledgebases, the Sinica BOW model can also be applied to encode and represent a particular knowledge system, such as Tang civilization. This application will allow comparative studies of a historical conceptual system with our modern conceptual system. Our pilot study on the 300 Tang Poems is

reported here. The segmented and classified lexicon of the 300 Tang Poems (Chang and Luo 1999) serve as the basis of this study. Three domain ontologies are constructed and studied: animals, plants, and artifacts. Each domain is mapped to the SUMO/BOW structure. The resultant ontological representation is taken as a slice of the knowledge structure of Tang civilization. For instance, from the ontology of animals of Tang 300 (see attached file), we reach some broad generalizations about the familiar fauna of Tang. With further examination, we also found a fascination with flying in Tang is confirmed by the poetsí choice of poetic animals.
In sum, we argue that the Sinica BOW model will not only be a useful resource but also a productive model for the construction of a knowledgebase that will greatly facilitate our understanding of Tang civilization.

1 Background

The construction of an ontology from a knowledge background which is substantially different from ours can be challenging yet rewarding. We will refer to this type of ontology as "Non-Standard Ontology" for lack of better terms. Work on non-standard ontology presents a dilemma. On one hand, the structure of knowledge is often neither explicated nor represented before the non-standard ontology is constructed. On the other hand, to construct such an ontology, one needs to start with at least some pre-defined terms and conceptual taxonomy, which is in practice a small (upper) ontology. For historical ontologies, it is very rare to find a synchronous ontology from the same period, such as Wilkins (1668). In this case, the structure of the synchronous ontology can be adopted and mapped to a modern system for study. However, for the knowledge domains with no existing ontologies available, the greatest challenge also underlines the greatest potential to gain new knowledge. For instance, seventh century Chinese does not have the same scientific knowledge or the philosophical tradition that the current academic world holds to be common. Hence, even though there is much knowledge to be gained, there is also very little to fall back to as the working hypothesis. We will show in this paper how such dilemma can be resolved with successful integration of lexical resources and upper ontology.

The target ontology of this study is the ontology of the Tang dynasty (618-907AD). In this pilot study, we work with the text of the collection of the Tang 300 Poems. We adopt SUMO as our upper ontology. The lexical resources used include the domain lexica extracted from the text and the English-Chinese bilingual wordnet system Sinica BOW.

2. Sinica BOW: lexicon based bilingual knowledgebase

Lexicons can perform the bridging function between documents and conceptual categorisation. This position is motivated by both language engineering concerns as well as psychological felicity. In addition, when the issues and needs of multi-linguality are taken into consideration, it becomes obvious that the lexicon is the only level where generalizations as well as variations across different languages can be captured efficiently and comprehensively. In this demo, we will show our work on integrating multiple lexical

resources with ontology such that the linguistic-to-conceptual representation and language-to-language gaps can be bridged simultaneously.

The ***Sinica BOW*** (Academia Sinica Bilingual Ontological Wordnet) is intended as a linguistic infrastructure for knowledge representation and knowledge engineering. It is built upon the relation-based structure of WordNet. On one hand, a bilingual wordnet is constructed with the crucial design feature of treating bilingual translation correspondences as lexical semantic relations (Huang et al. 2003). On the other hand, SUMO (Suggested Upper Merged Ontology) is adopted as the shared system of conceptual categorization (Niles and Pease 2001). SUMO is also one of the first conceptual categorization systems to be mapped to an English lexicon (Niles and Pease 2003). Since SUMO is mapped to WordNet 1.6 (and most recently to WordNet 2.0), the English WordNet has become the cornerstone for linking across languages and between a language and its conceptual system. In addition, domain tags are assigned to lemmas when necessary in order to ensure domain inter-operability.

By the combination of ontology and wordnet, we hope that Sinica BOW will 1) give each linguistic form a rigorous conceptual location, 2) clarify the relation between conceptual classification and linguistic instantiation, and 3) facilitate genuine cross-lingual access of knowledge.

The Sinica BOW allows lexical searches in either language to return ontological information (in either language). Searches on Sinica BOW can return the following information: Sense-based English-Chinese translation equivalency, English word-sense-based ontology and inference, Chinese word-based ontology and inference, Word-sense-based domain specification (under construction).

In addition to the integration of Wordnet and ontology, it is also an important goal of Sinica BOW to integrate lexical resources. Sinica BOWís design is lemma-driven. A lexical database of word forms is first compiled by integrating multiple lexical resources. This becomes the central database for lexical management for Sinica BOW. Making use of this lexical database, a lexical search may link to either the main BOW knowledgebase or any of the corresponding entries in an online lexicon.

2.1. The Multilingual and Cross-Domain Properties of (Semantic) Relations

In addition to relying on lemmas as retrieval keys, a crucial step in establishing synergy between language and knowledge resources is to identify the conceptual atoms that apply equally effectively to knowledge and language resources. Lexical semantic relations are exactly such a set of atoms. Sinica BOW implements this idea by encoding the lexical semantic relations between English-Chinese translation equivalent pairs. In addition to more precisely describing the relationship between two translation equivalents, this also allows better cross-lingual inferences. Explicitly allowing lexical semantic relations to be coded cross-lingually also will facilitate the transferring to a structured set of tree relations from one language to the other.

2.2. Taking Advantages of Lexical Structure*s*

In addition to the integration of bilingual WordNet and SUMO, Sinica BOW also integrates the rich structural information of the integrated lexical resources. Glyph,

phonological, and morphological structures can all be used to help access the ontological wordnet. This work has implications far beyond being convenient search tools. It is often claimed that the glyph composition (e.g. radicals) in Chinese has its semantic base. This can also be said about the morphological composition (and to a much lesser degree, phonological composition). In other words, the integration allows us to study the possible links between these lexical structures and conceptual classifications.

2.3. Conclusion

Integrating and interpreting information from multiple and varying sources will be the main challenge for information processing for the current generation. Taking lexicon as the bridging knowledgebase and ontology as the overall knowledge structure seems to be a logical choice. Integrating the two resources with multilingual capacity will add to the versatility and open new possibilities.

3. Resources and Structure of Sinica BOW

Conceptual structure and lexical access are two essential elements of human knowledge. Bilingual representation of both conceptual structure and lexical information will enable language independent knowledge processing. In this paper, we introduce a new type of integrated language resources: Bilingual Ontological Wordnet. The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) was constructed in 2003. We argue that such combination of ontology and wordnet will 1) give each linguistic form a rigorous conceptual location, 2) clarify the relation between the conceptual classification and its linguistic instantiation, and 3) facilitate genuine cross-lingual access of knowledge.

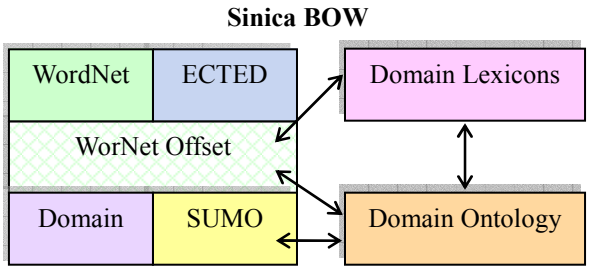
The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED), and SUMO (Suggested Upper Merged Ontology).

WordNet is a lexical knowledgebase for English language that was created at Cognitive Science Laboratory of Princeton University in 1990 (Fellbaum 1998). Its content is divided into four categories based on psycholinguistic principles: nouns, verbs, adjectives and adverbs. WordNet organizes the lexical information according to word meaning and each synset groups together a set of lemmas sharing the same sense. In addition, WordNet is a semantic network linking synsets with lexical semantic relations. WordNet is widely used in Natural Language Processing applications and linguistic research. The most updated version of WordNet is WordNet 2.0. We adopted WordNet 1.6., the version which is used by most applications so far.

ECTED was constructed at Academia Sinica as a crucial step towards bootstrapping a Chinese wordnet with English WordNet (Huang et al. 2002, Huang et al. 2003). The translation equivalence database was hand-crafted by the WordNet team at CKIP, Academia Sinica. First, all possible Chinese translations of an English synset word (from WN 1.6.) are extracted from several available online bilingual (EC or CE) resources. These translation candidates were then checked by a team of translators with near-native bilingual ability. For each of the 99,642 English synsets, the translator selected the three most appropriate translation equivalents whenever possible. The translation equivalences were defaulted to lexicalized words, rather than

descriptive phrases, whenever possible. The translation equivalences were then manually verified. Note that after the first round of translation, there were about 5% of the lemmas whose Chinese translation can neither be found in our bilingual resources nor be filled by the translators. We spent another 2 person-year consulting various special dictionaries to fill in the gaps.

Figure 1: The resource and structure of Sinica BOW:



SUMO is a upper ontology constructed by the IEEE Standard Upper Ontology Working Group and maintained at Teknowledge Corporation. SUMO contains roughly 1,000 conceptual nodes for knowledge representation. It can be applied to automated reasoning, information retrieval and inter-operability in E-commerce, education and NLP tasks. Niles & Pease (2003) mapped synsets of WordNet and concept of SUMO in three relations: synonymy, hypernymy and instantiation. For instance, the synset "animal" (a living organism characterized by voluntary movement) in WordNet is synonymous with the SUMO concept of "Animal". In "bank" (a financial institution that accepts deposits and channels the money into lending activities) this case, bank is a corporation that is a hypernym of the associated synset. President of the United States (the office of the US head of state) is an instantiation of "position" concept. Through the linking and the interface available at the SUMO website (http://ontology.teknowledge.com), each English lemma can be mapped to a SUMO ontology node.

The three above resources were originally linked in two pairs: WordNet 1.6 was mapped to SUMO by Niles and Pease. ECTED maps English synsets in WordNet to Chinese lexical equivalents, which encodes both equivalent pairs and their semantic relations (Huang et al. 2003). WordNet synsets thus became the natural mediation for our integration work. Thus, with the integration of these three key resources, Sinica BOW can function both as an English-Chinese bilingual wordnet and a bilingual lexical access to SUMO. In other words, Sinica BOW allows a 2x2x2 query design, where a query could be in either Chinese or English, either in lexical lemmas of SUMO terms, and the query target can either be the wordnet content or the SUMO ontology.

The design of Sinica BOW has an additional domain information layer, as shown in figure 1. The domain information will be represented by a set of Domain Lexico-Taxonomy (DLT, Huang, Li, and Hong 2004). In this design, our main concern is domain inter-operability. It can be safely assumed that domain exclusive words (i.e. lemma-sense pairs) are recorded only in domain lexica, hence there will be no ambiguity and no inter-operability issues. We concentrate instead on the lexical items that intersect with the general lexicon. On one hand, since these are the lemmas that may occur in

more than one domains with one or more different meanings, domain specification would help resolving the ambiguity. On the other hand, these general lemmas with domain applicability can be effective signatures for the applicable domains. The real challenge to domain inter-operability involves the ëunknowní domains where no comprehensive domain lexica/corpora are available. We argue that this problem can be greatly ameliorated by tagging the general lexicon with possible domain tags. When domain tags are assigned to lemmas whenever possible, the general lexicon will contain substantial partial domain lexica. Although we cannot expect to construct full-scale domain lexica within the general lexicon, these domain-tagged lexical items serve as a scalable basis for future bootstrapping for domain lexica.

4. Presentational Versatility

Sinica BOW allows versatile access and facilitates a combination of lexical semantic and ontological information. The versatility is built in with bilinguality, and lemma-based merging of multiple language sources.

The versatility and combinatory presentation is crucial to the presentation of a knowledge system.

4.1. Lexicon-driven Access

Since the main goal of Sinica BOW concerns knowledge representation, the lemma based or conceptual node based query results are directed linked to the full knowledgebase and expandable. The Sinica BOW access is lexicon-driven. Each query returns a structured lexical entry, presented as a tree-structured menu. A keyword query returns with a menu arranged according to word senses, as shown in Figure 2. The top level information returned including POS, usage ranking, and cross-reference links. In addition to wordnet information, cross-references to up to five resources are pre-compiled for either language. For an English word, the main resource is of course the bilingual wordnet information that our team constructed. Major outside references are listed for quick hyperlink. These include corpora and both EC and CE dictionaries. For Chinese, the main resource is again our bilingual wordnet. In addition, links are established to Sinica Corpus, to Wen-Land (a learnerís Lexical KnowledgeNet), and to online monolingual and bilingual dictionaries. In addition to online access of multiple sources information, each lemmaís distribution in these resources is also a good indicator of its usage level.

⊞ 詞義(Sense)4: 魚兒

領域 Domain	一般(General) 建議-fish Sense 4-的領域值
POS 詞類	名詞(Noun)
解釋 Explanation	any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills
翻譯 Translation	魚兒, 魚
同義詞集 Synset	fish
（整體）部件詞 Part meronym	milt, tail_fin, fishbone, fish_scale, fin, roe, caudal_fin, lateral_line_organ, lateral_line
上位詞 Hypernym	aquatic_vertebrate
下位詞 Hyponym	food_fish, game_fish, rough_fish, cartilaginous_fish, chondrichthian, bony_fish, mouthbreeder
（成員）群體詞 Member holonym	shoal, Pisces, school
SUMO	同 fish:Fish(魚類)

Figure 2: A sample lemma query result of Sinica BOW

The access to the ontology and the domain taxonomy are also lexicon-driven. That is, in addition to using the pre-defined ontology or domain terms (in either English or Chinese), a query based on a lexical term is also possible.

For SUMO, it will return a node where the word appears in. It can also be achieved by looking up the ontological or domain node the word belongs to.

One last but critical feature of the lexicon-driven access is the possibility to re-start a query with any lexical node. When expansion reaches at the leave node and results in a new word, clicking on the word is equivalent to start a new keyword search.

4.2. Multiple Knowledge Source

Sinica BOW preserves the logical structure of both WordNet and SUMO ontology yet links them together to allow direct accesses to the merged resources. This is shown in Figure 2. In a wordnet search, the return includes an expandable list of the complete bilingual wordnet fields. The fields are listed under each sense and include: POS, synset, sense explanation, translation, and list of lexical semantic relations. In addition, we add the domain information, translation equivalents, and link to the corresponding SUMO node. Each field is expandable to present the database content. For instance, Figure 2 shows the query return for the lemma "fish", with the Part_Meronym and Holonym of sense 4 expanded. The field of domain and SUMO will lead directly to the corresponding node in the domain taxonomy of the ontology and allow further exploration. For instance, the menu item of the mapped SUMO node links to the SUMO representation, as well browsing of the SUMO ontology and axioms.

Two more aspects of versatility can be achieved through the use of higher level linguistic generalizations and the use of domain taxonomy to organize information. These will be discussed in more details in the next section.

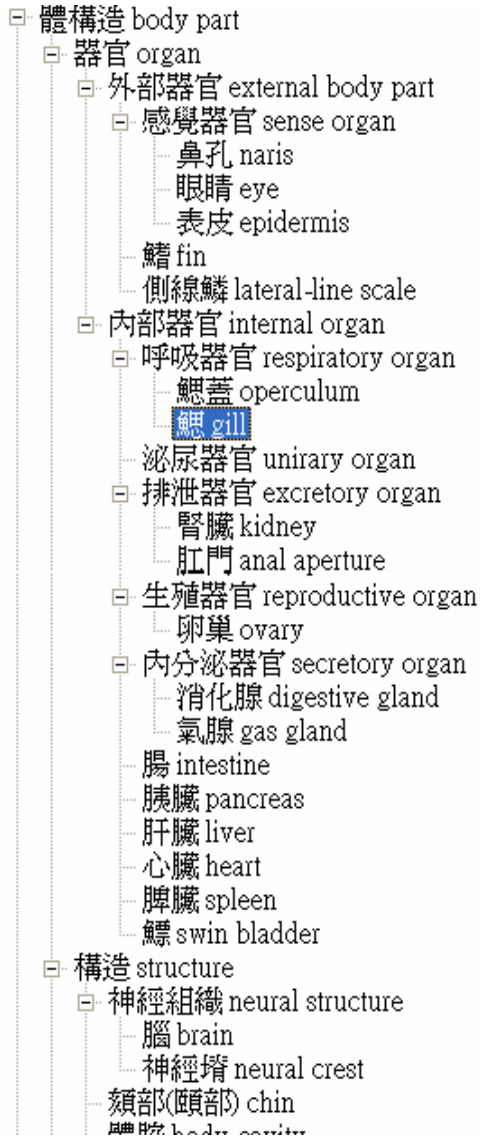


Figure 3: A sample domain ontology: Fish

5. Higher Level Generalizations

Linguistic as well as resources structures are utilized in Sinica BOW to facilitate formation of generalizations as well as to assist queries where the user is not sure of the precise lemma form. The non-lexical access includes

alphabetical (for English), prefix (for Chinese, including root compounds), suffix (for Chinese, including root compounds)), POS, frequency, domain, SUMO concepts, as well as a combination of the above conditions. With this additional level of resource integration, generalizations such as the semantic correlation of senses and morphological heads can be easily reached.

Domain taxonomy can also be utilized to organize and access information. Our Domain Lexico-Taxonomy approach attempts to assign a domain tag to a word whenever applicable. We also encourage users of SUMO to feedback with their own domain use of lexical items because domain specifications can not be covered by any single knowledge source. Hence we BOW contains rich domain information. Hence we also allow structured access to the Sinica BOW knowledge content by specifying a node on the domain taxonomy. This feature enables quick extraction and checking of a domain lexicon.

6. Domain Ontology

One of the most immediate and perhaps most powerful application of Sinica BOW is perhaps the construction of domain specific ontologies. This will be a crucial step towards providing a feasible infrastructure to implement web-wide specific ontologies, as required by the vision of Semantic Web. It is also a critical test to see if the upper ontology approach is really applicable to a wide range and diversity of knowledge domains. And lastly, for Sinica BOW, it provides a test ground for us to show that the combination of bilingual wordnet and ontology does provide a better environment for knowledge processing.

Two first attempts have been carried out. The first is a small fish domain ontology projected from the FishBase terms. This is mapped using Sinica BOW. Part of the ontology is shown in Figure 3. We would like to explore the possibility of using this domain ontology for non-expert to extract expert knowledge from the FishBase in the future.

The second attempt, reported in Huang et al. (2004), involves the Shakespearean-garden approach to domain ontology. In this approach, we collect domain lexicon from a target collection of texts (Tang poems in this case), and map them to the SUMO ontology. This approach allows us to examine the knowledge and/or experience of a specific domain as reflect in that collection of texts. This could be personal, historical, regional etc. This approach allows us to make generalizations based on the full knowledge structure, not just one lexical incident. For instance, we were able to confirm the Tang civilizationís fascination with flying by looking at the dominance of animal references in the texts.

7. The Shakespearean-garden Approach Toward Non-standard Ontology

We propose a Shakespearean-garden approach to the construction of non-standard ontology. This approach is both lexicon-based and domain-driven. A Shakespearean garden collects and grows all plants referred to in Shakespearean texts. The purpose of a Shakespearean garden is to replicate the botanic knowledge and flora experience of Shakespearean England. A Shakespearean garden works because we can reasonably assume that the plants we collect now are by and large identical to the Shakespearean plants and have the same functions. Similarly, when constructing a non-standard ontology, we propose to start with concrete sub-domains. A

chosen domain must have two properties: that it plays roughly equivalent roles in the knowledge backgrounds of the target ontology and the reference ontology (i.e. our contemporary ontology); and that it is empirically verifiable with lexical resources supporting the target ontology. Even though the Shakespearean-garden approach does not guarantee a complete ontology, it will lead to very reliable domain ontologies. When there is sufficient data and knowledge collected, these domain ontologies can be further linked to approach a complete ontology of the target knowledge domain.

Our approach requires a shared upper ontology as the anchor for bootstrapping and for comparative studies. We assume that when two knowledge systems are studied, there will be no meaningful comparison unless both of them can be put in the same representational framework. In the current work, we adopt SUMO (Suggested Upper Merged Ontology, Niles and Pease 2003) as the framework for ontological representations. SUMO was constructed with the explicit goal to serve as the upper ontology of varying knowledge domains by the IEEEís suggested upper ontology workgroup. In other words, SUMO is supposed to be versatile and has robust coverage of general concepts used by different ontologies. Since SUMO is attested with many contemporary knowledge domains, it offers a good foundation for our comparative study of non-standard ontology. In addition, our application to a temporally and culturally far removed knowledge source offers a genuine challenge to the robustness of SUMO. Lastly, as an upper ontology, SUMO avoids elaboration of lower level nodes. Hence there is only a very low probability that it will run into contradictions with the expanded nodes of a non-standard ontology.

While an upper ontology is adopted as the anchor for domain ontology construction, such an upper ontology may not contain all the finer-grained concepts necessary to fully represent the chosen domain. Hence, we propose to use Wordnet to supplement the knowledge. Wordnet as a lexical knowledgebase provides the natural interface between the domain lexica and SUMO (Niles and Pease 2003). In addition, for concepts not explicitly represented in the upper ontology, wordnet lexical semantic relations can be used to construct a conceptual taxonomy.

All the lexical and knowledge resources required for this approach are already integrated in Sinica BOW (Academia Sinica Bilingual Ontological WordNet, Huang et al. 2004). Hence we use Sinica BOW as the primary referential knowledgebase in this study. Sinica BOW integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED, Huang et al. 2003), and SUMO. Referring to Sinica BOW has three advantages. First, it allows access to both lexical semantic relation in WordNet and conceptual taxonomy in SUMO. Second, it allows lexical search in either Chinese or English. Third, it allows research information to be represented in either Chinese or English.

8. Mapping Lexical Data to Ontology

8.1. Preparing the Lexical Resources

Tang civilization (618-907AD) was one of the most vibrant periods of Chinese civilization. It welcomed and integrated elements from many of the neighboring non-Han civilizations. In turn, Tang civilization was also venerated and imitated by neighboring countries. The Japanese civilization, for instance, borrowed

generously from Tang, including the kanji writing system. It is not an exaggeration to claim that the classical roots of Japanese civilization are actually Tang civilization. Hence, the ontology of the Tang dynasty has far more implications than being an ontology of a long-gone historical period. It may shed light on how heterogeneous knowledge systems integrate, as well as how a borrowed knowledge system develops in the new cultural background.

As a pilot of the main study of constructing an ontology based on the more than 10 millions characters in textual archives from the Tang Dynasty, we construct an ontology based on the famous anthology of The 300 Tang Poems. The text of the 300 Tang Poems contains slightly more than 15,000 characters. This is one of the most important and popular collections of Chinese literature. Its importance far out-weights its relative small size. In addition, since it is poetry, the conceptual density, as represented by the lexical types contained, is high. In this pilot study, the words and classification of words in the text are hand-tagged. The choice of manual tagging is made because our tagger is not tested for domain classification, even though it performs the task of pos tagging very well. The relatively small size of the text also allows manual work to be done efficiently. The highly reliable result will serve as valuable training data for future automatic tagging classification. There is already a classical Chinese tokenizer combining segmentation and tagging available from Academia Sinica. This tokenization program, adopting the basic design of Chen and Liu (1992), is very robust and performed well in the first SigHAN Chinese segmentation bakeoff in 2003. It has also successfully segmented over 5 million words of classical Chinese texts for the language archives project at Academia Sinica.

Three sub-lexicons from the Tang 300 Poems were extracted for domain ontology construction: animals, plants, and artifacts. A total of 176 words were assigned to the three domain lexica: The animals lexicon contains 64 words; the plants lexicon contains 59 words; and the artifacts lexicon contains 53 words. The result from the animal and plant domains will be reported in this paper. These domains are chosen because their meanings are referential and rich. Since they are referential, it is more likely to uniquely determine the meaning of each term. On the other hand, these are familiar terms and important poetic devices used to invoke empathy or express feelings.

The second step in the preparation of the lexical resources for ontology-building is the identification of the appropriate sense of each word for the target knowledge domain. There are two issues involved here. First, as most words are assigned more than one senses in wordnet, we need to identify the correct sense. Second, as these words are used over 11 hundred years ago, some meanings may have become obscure or changed. We need to identify the intended meaning. A batch query on these 176 words was sent to Sinica BOW. Of the 176 words, only 100 words found complete matching entries in the Chinese part of the bilingual wordnet. We then expand the query to include words that share the initial or ending characters. The expanded query still left 24 words with no possible matches in the current version of BOW. These 24 words were later assigned correct translation and meaning with manual dictionary lookup. For words with direct sense assignment from WordNet, the link form BOW to SUMO

ontology is utilized. When a sense does not belong to the target knowledge domain, it is discarded. The senses that belong to the target domain by SUMO assignment is kept for next step. Even though there were in average 2.18 senses assigned for each word, the domain requirement quickly reduced the number of possible senses to close to one.

It is important to notice that expertise knowledge is crucial in the identification of word senses when dealing with a non-standard knowledge domain. A good example is the word *mei2*, with grass radical found in the Tang poems. Its dominant sense in contemporary Chinese equals to berry, as in strawberry "*cao3mei2*". However, further investigation showed that such sensed did not exist in Tang dynasty. The word refers to a kind of moss instead. In other words, although the Chinese character composition reinforces its position in the plants domain, its actual reference cannot be reliably determined by using standard lexical knowledge.

Expertise knowledge and manual editing is also crucial for the words that do not find direct match in Sinica BOW. For example, *hu2jia1* is a particular musical instrument that was first invented and played by the Tartar people and no longer commonly used. Hence its lack of an equivalent in the English language is not surprising. To solve this problem, we consult similar senses from Wordnet. Since *hu2jia2* is a kind of tubular wind instrument, we considered it to be a kind of pipe, which does occur in WordNet and is linked to SUMO.

8.2. Constructing Domain Ontology

Once each lexical item is assigned a unique correct Chinese sense and its corresponding English synset, it can be mapped through Sinica BOW to a SUMO conceptual node. When there is no exact match, lexical semantic relations from WordNet are consulted to establish relation between a lexical item and SUMO. For lexical items that are thus assigned to an appropriate SUMO node, the construction of the domain ontology is as simple as connecting two dots. This is largely the case for the animals ontology (Figure 4).

On the other hand, SUMO as an upper ontology does not necessarily offers sufficient knowledge structure for all domains. For instance, although plants can be considered to be equally salient as animals conceptually, SUMO only gives the very rough-grained classification of FloweringPlant and NonFloweringPlant. Hence we need to use the lexical semantic relations from WordNet to construct the hierarchical conceptual network, i.e. the proposed domain ontology. In this case, we cannot simply copy and connect the relations. Since WordNetís main goal is to record all cognitively relevant semantic relations, not all relations can fit in a rigorous conceptual classification and inference system. Hence, after bootstrapping with all WordNet synsets and relations marked, an important step is to prune the resultant tree for both inconsistency and redundancy. The plants ontology in Figure 5 is the wordnet-based ontology after extensive pruning.

In establishing the link between a sense and a ontology node, it is important to notice that the SUMO-WordNet link is established with the contemporary background knowledge of the English speaker world. Hence it is likely to find that a non-standard ontology based on a different system will require a totally different conceptual assignment. An instance of such mismatches involves *mou2hu2*, which is a kind of silk flag. A flag, according to both the literary context and the assigned lexical

sense, should be a piece of artifact, solid and substantial. However, the SUMO-WordNet link that Sinica Bow follows mapped it to the conceptual node of "Icon". This may be appropriate when a flag is used in signing, but not appropriate in the Chinese context. Hence we simply correct the link and assign it to artifact.

What is more interesting in terms of linguistic use involves words that seem to carry the same meaning, while involves fundamentally different conceptualization. The difference in conceptualization requires assignment to a different ontological location. One such example is *dai4mei4*, which is given the sense of "a beaded sea turtle", and seems to be a straightforward case of a kind of animal. However, when we refer to the context, the sentence actually refers to "a beam inlaid with *dai4mai4*". In other words, it refers to the materials used in decorating a building. It is the shell of the turtle that has been ground and polished like a piece of jade. It is also interesting to note the fact that these two characters used have a jade radical, rather than an animal or fish radical. Both the context and the written form suggest that the sense being used here is the material, and there is no evidence suggesting that Tang people know that the *dai4mei4* material comes from a turtle. Hence this word is not included in the animals ontology.

On the other hand, when metonymy is used, it is often possible to argue that the original sense is invoked. An example in our study is *shuang1li2*, double-carp, which refers to a letter since letters are traditionally sent in a wood box with two carps carved on top. In this case, even though the actual reference is not the animal, but the lexical metonymy necessarily involve the image of the fish. Hence we consider the concept of carp is used, and hence justifying our including carp as an attested case for the animals ontology for Tang.

9. Result and Discussions

The result of this pilot study will include three semi-automatically constructed sub-ontologies: animal, plant, and artifact. The first two are completed and will be discussed here. The top part of each ontology is mapped to SUMO. The lower part of each ontology is extended using WordNet relations. These ontologies as well as the attached lexical terms will have Chinese-English bilingual representation.

The first generalizations that can be obtained are from the distribution of these domain terms in the texts. The total frequency of these three domains ranges from 1.65% to 1.89%. These are relatively high compared to a balanced corpus. In a balanced corpus, the top 20 animal or plant domain terms comprise of less than 1%.

The second generalizations can be made from the distribution among the different terms within the domain. Among animal concepts, the total frequency of birds is over 38%, and hoofed mammals over 30%. These two kinds each far exceed all the other eight kinds of animals combined. This fact should have implications on either the fauna of Tang, or the poetic choice of images. Even more striking is the fact that of all plants, flowering plants consist of over 95% of the instances in the texts. This fact should not be surprising because of the strong poetic image that a flower presents.

After the sub-ontologies are constructed, comparative studies of the Tang ontological structure with our contemporary ontology (based on SUMO) will be conducted. For

instance, we found that among the order of mammals, the families of marsupials and marine mammals are missing. The absence of marsupials is expected since it is a fact of science history that they were discovered much later. The absence of marine mammals may point to the fact that the Tang civilization is mainly land-based. In addition, we also found two interesting facts in other branches. First, almost all invertebrates that are documented are (winged) insects. And among the non-mammal vertebrates, with only less than 5 exceptions, all documented lexical items refer to bird. A possible explanation of the idiosyncrasy is the Tang civilizationís fascination with flying. We know as a fact that flying is a recurring theme in paintings from this period, and occur in poetry too.

The plants ontology of Tang offers a good test case of how to bootstrap an ontology with lexical knowledgebases such as wordnets. We showed that when the lexical resource contains sense and lexical semantic relations information, it is possible to use the information to bootstrap a domain ontology. The crucial challenge here is how to turn the set of pair-wise and lexicon-driven relations to a taxonomical hierarchy. An issue that will recur is how to deal with same level nodes that are classified and assigned with diagonal criteria. One such example is the classification of plants in Figure 5. FloweringPlants and HerbaceousPlants and AcquaticPlants create partially overlapping classes. These are all linguistically and cognitively motivated and cannot subsume each other. Given the fact that even an upper ontology like SUMO acknowledes such human cognitive facts and allows multiple inheritance, there is still reservations that an ontology can quickly become non-trackable if no constraints are put on such cross-classification. This is an issue that merits in-depth formal and theoretical deliberation.

10. Conclusion

In this current study, we propose the Shakespearean-garden approach to the construction of non-standard ontology. We showed with a pilot study that such an approach is feasible, especially when supported by the right combination of lexical knowledge sources and upper ontology. In addition, we showed that the constructed sub-ontology allows us to have a comprehensive view of the knowledge system of a civilization that no longer exists. Such a representation will offer a unique opportunity to study how their world differs from ours and how they view the world differently from us.

A natural extension of the current work is to try to piece these sub-ontologies together to form a skeletal ontology for the Tang dynasty. In order to carry out this full-scale work, we have already started the design and construction of automatic tools to construct domain ontology based on domain lexicons and SUMO. This will integrate the knowledge we gain from the current work as well as modules from existing systems, such as Sigma system constructed by Adam Pease. Such a working environment will facilitate the ultimate goal of the Shakespearean-garden approach. In addition, we will also try to apply the simultaneous bilingual mapping approach to construct a modern domain. Ultimately, we would like to see if it is still plausible to construct ontology based on a shared upper ontology even if the background knowledge systems are drastically different.

The current work on the domain knowledge of Tang civilization willl also

provide solid foundation for future work on metaphor. Based on Lakoffís contemporary theory of metaphor, Ahrens et al. (2003) shows that the crucial step in predicting and explanation of the use of linguistic metaphors lies in capturing the rules governing the mapping between source domain and target domain knowledge. For the historical poetic work such as Tang poetry, an additional challenge to the study of metaphor would be the precise characterization of the source domain knowledge. Our non-standard ontology can be viewed as the foundational work defining source domain knowledge in Tang poetry. With the source domain knowledge described, we will be able to develop in-depth study of Tang poetic metaphors in the future.

Lastly, the issue regarding the relation between a wordnet and an ontology is also touched upon. In the Shakespearean-garden approach, it is crucial that the specific domain lexicon can be obtained and annotated with correct lexical semantic information. However, how can lexical semantic relations be best used in an ontological study remains a challenging and promising issue.

Online Resources

Sinica BOW: http://BOW.sinica.edu.tw/
SUMO: http://ontology.teknowledge.com/
WordNet: http://www.cogsci.princeton.edu/~wn/
Tender Lyrics-The 300 Tang Poems (in Chinese) http://cls.admin.yzu.edu.tw/300/HOME.HTM
CKIP Segmentation and Tagging Program
http://corpus.ling.sinica.edu.tw/project/LanguageArchive/lc_index.html

Reference

Ahrens, Kathleen, Chu-Ren Huang, and Siaw-Fong Chung. 2003. Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles. Presented at the Workshop on Lexicon and Figurative Language. An ACL2003 Workshop. July 11, Sapporo, Japan.

Chang, Ru-Yng and Feng-ju Luo. 1999. Cross-platform Web-bases Learning Systemó the construction of Tender Lyrics-The 300 Tang Poems (in Chinese). Presented at 1999 Taiwan Symposium on Taiwan Academic network. Kaohsiung.

Chen, K.-J. and S.-H. Liu. 1992. Word Identificaiton for Chinese Sentences. Proceedings of COLING92. 501-505.

Fellbaum, Christine. Ed. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Huang, Chu-Ren, Ru-Yng Chang, and Shiang-bin Li. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. To be presented at the LREC2004 conference. May26-28. Lisbon.

Huang, Chu-Ren, Li, Xiang-Bing, Hong, Jia-Fei. (2004). Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing. Proceedings of the Asian Symposium on Natural Language Processing to Overcome Language Barriers. March 25-26, 2004. Hainan Island.

Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. (2004). Sinica BOW and 300 Tang Poems: An overview of a bilingual ontological wordnet and its application to a small ontology of Tang poetry. Presented at the Workshop on Possibilities of a Knowledgebase of Tang Civilization. Institute for Research in Humanities, Kyoto University. February 20-21.

Huang, Chu-Ren, Elanna I.J. Tseng, Dylan B.S. Tsai, & Brian Murphy. (2003). Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations. Language and Linguistics. 4(3), 509--532.

Huang, Chu-Ren. Elanna I.J. Tseng & Dylan B.S. Tsai. (2002). Translating Lexical Semantic Relations: The first step towards multilingual Wordnets. In Proceedings of the COLING2002 workshop: SemaNet: Building and Using Semantic Networks. Taipei, Taiwan.

Niles, I. & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003), Las Vegas, Nevada.

Niles, I., & Pease, A., (2001). Toward a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Ogunquit, Maine.

Wilkins, J. (1668). An Essay Towards a Real Character, and a Philosophical Language. Reprinted in 2002. Thoemme Press.

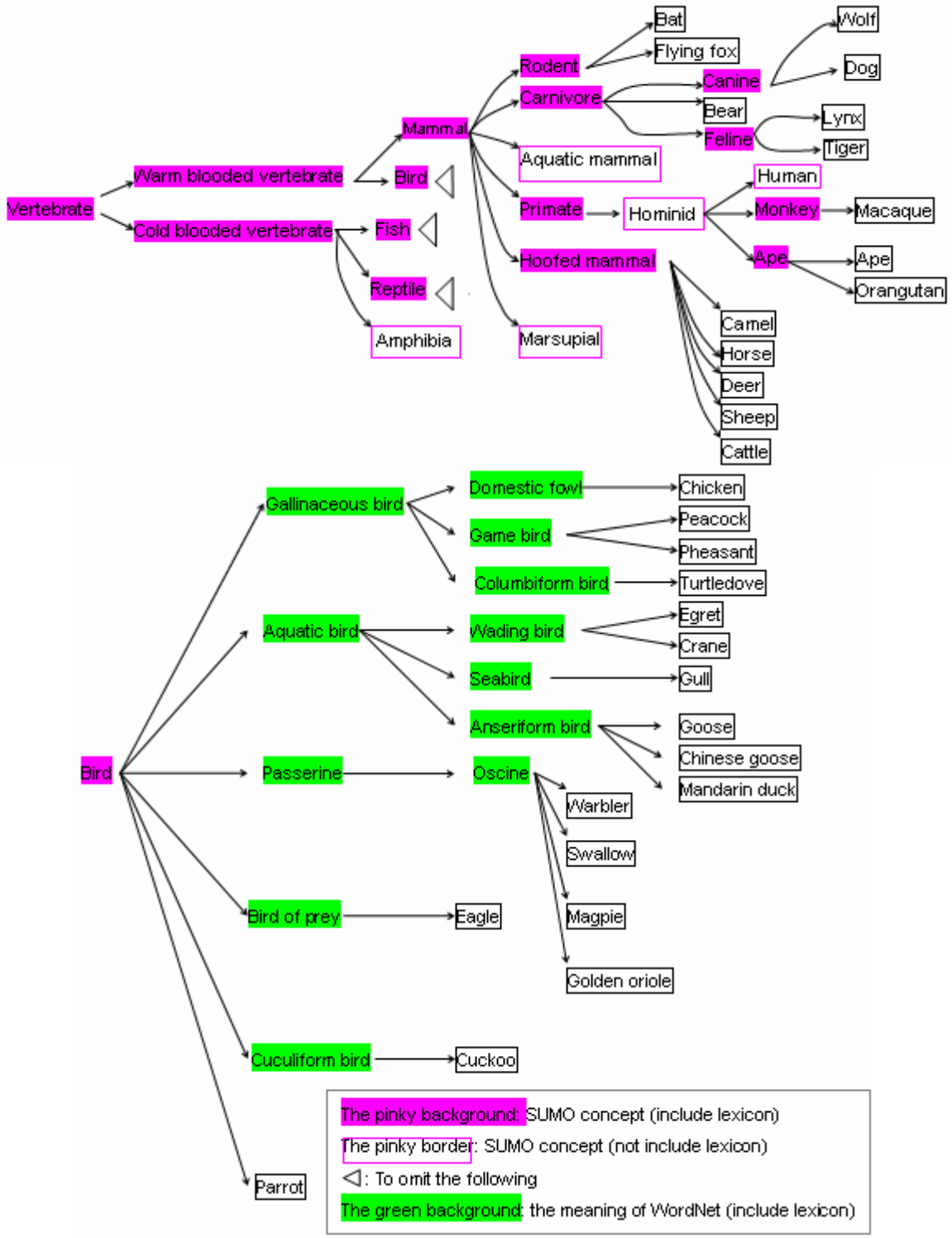


Figure 4: Tang Animals Ontology

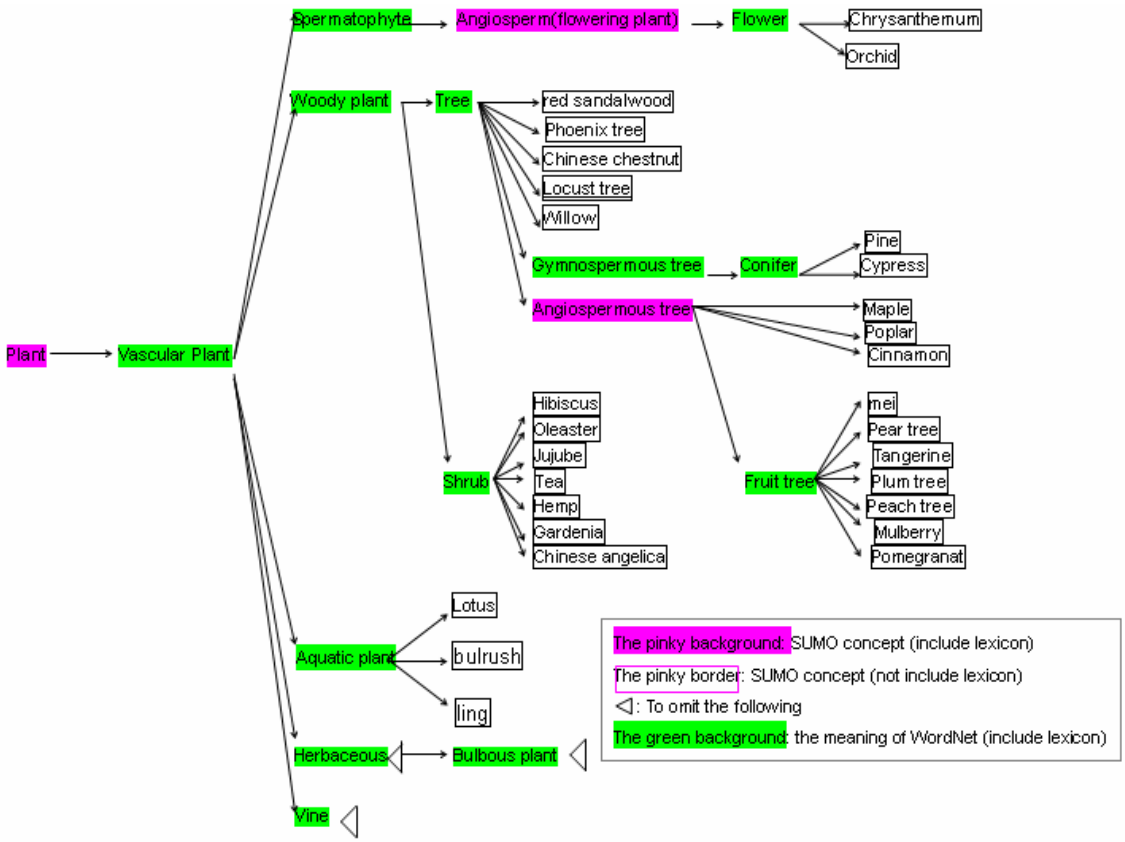


Figure 5: Tang Plants Ontology