專 題 演 講

演 講 者：項潔 教授(台灣大學資訊工程系)

演講題目：Building digital archive systems for historians

演講摘要：

The maturity of digitization technologies has provided historians with an unprecedented amount of historical archives in digital form. Generally speaking, however, the systems that have been built for using these digital archiveshave not taken into consideration the special needs of historical research.

Conventional wisdom dictates that a retrieval system should yield high precision for a casual user while provide high recall for scholarly use. In order to ensure high precision, a retrieval system assumes that documents are independent from (or even competing with) each other so that those deemed more relevant to the query will emerge on top of the resulting list. High recall is then achieved through enhancements such as query refinement or a thesaurus to get as many potentially relevant documents as possible.

We noticed from our interaction with historians that they consider high recall desirable but not crucial. While there is always the concern of missing something important, many historians are also worried about being overwhelmed by the large number of query returns resulted from high recall. A more fundamental issue that had been overlooked is that historians usually consider documents as related, not independent as assumed by document retrieval systems. Indeed, a historian rarely looks at a single document alone, but rather a group of documents and searches for properties that the documents collectively possess. In this sense historical studies is a research of context: context among documents; context between documents and the intangible societal, cultural and historical factors; or even context observed from missing patterns. To respond to this challenge, we suggest that a digital archives system for scholarly use should provide the user with the collective meanings of a query result set, and not just high recall alone.

The underlying design philosophy of the digital archive systems that we describe in this talk is to treat a set of query returns as a meaningful sub-collection. Instead of ranking the documents (as is done in the precision/recall model), the proposed system returns a list of documents when issued a query, together with a choice of textual contexts of the returned sub-collection as well as

visual mechanisms to observe, explain, and refine the contexts. In other words, in addition to providing search and retrieve, the system also allows the user to observe, analyze, explore, and discover textual contexts among the retrieved documents. The contexts and their visualization are expressed in such a way that they can be interpreted by the user. Thus, the historian can observe and explore relationships among documents in an unprecedented scale and ease. Through exploring and analyzing the contexts, the historian can also discover research issues that had not been noticed before. What we are advocating, then, is the system as a digital platform that evolves from providing information to discovering research problems. We emphasize that the system is an observation environment and is not meant to replace the historian's decision-making process. A digital system can provide textual context: observable textual contexts within the documents, but not contexts that associate the documents with the external world. The interpretations of the findings and the narratives are still within the scholar's realm.

In this talk we shall give a sampling of digital archive systems designed under this principle. Although the historical materials in our systems are mostly in Chinese and are text-based (mainly Taiwanese history, Chinese history, and Chinese classical writings), the principle is the same regardless of the language or nature of the archives. We shall demonstrate different aspects and features of the kind of "explainable textual context" that are advocated in this abstract. Some typical contexts are chronological distribution, geographical distribution, term frequencies and co-occurrences of people and places, appositional term analysis, and holistic comparisons of query returns. How they can be used for discovery will be discussed. Future directions will also be outlined.