史料整體分析工具之幕後──
介紹「臺灣歷史數位圖書館」的資料前置處理程序

陳詩沛[*]、項潔[**]、杜協昌[***]

摘要

「臺灣歷史數位圖書館」（THDL）是臺灣大學近年建置的一個大型全文史料數位圖書館，共含近八萬件的第一手臺灣史文獻，累積的全文字數已達約一億五千萬字。本文對 THDL 納入史料時進行的資料前置處理方式做說明。一般的數位圖書館在納入資料時，往往不假設這些資料之間有顯著的關聯，所以在納入分批建置的新資料時，除了重新製作檢索索引（re-index）之外，並不需做太多的前置作業。然而，爲了讓臺灣史研究者能更有效地利用與分析 THDL 中的大量史料，我們發展了一系列的史料分析工具，這些工具需要對史料作整體性的分析，所以 THDL 對資料的處理方式遠比一般的數位圖書館複雜，尤其在輸入一批新資料的時候，需要與現有資料作整體、精細的前置處理，以預先建立全體資料的分析資訊與資料之間的關聯性，讓研究者有能力深入大量的史料，進行分析與研究。這個前置作業程序也反映我們打造 THDL 的原則，即是一個好的數位圖書館不應僅是一個資料倉儲，而應是資料與分析工具之間無縫的結合。

關鍵詞：數位圖書館、數位典藏、臺灣史、資料處理、數位研究環境

[*] 臺灣大學資訊工程系博士候選人，E-mail: gail@turing.csie.ntu.edu.tw
[**] 臺灣大學資訊工程系教授，E-mail: hsiang@csie.ntu.edu.tw
[***]臺灣大學資訊工程系博士後研究員，E-mail: tu@turing.csie.ntu.edu.tw

The Document Processing Workflow of THDL, Taiwan History Digital Library

Szu-Pei Chen[*], Jieh Hsiang[**], Hsieh Chang Tu[***]

Abstract

In this paper we describe the mechanism for incrementally incorporating new collections into the Taiwan History Digital Library, a digital research environment for research in Taiwanese history. In a system such as the THDL, in which documents and tools are fully integrated, adding a new collection involves a lot more work than meets the eye. The challenge mainly comes from the need to provide a global view of the entire document set. Our mechanism ensures that a newly added collection is fully integrated with the existing ones, and that a global view of all the documents in the digital library is always preserved. It also reflects our philosophy that a digital library is not merely a warehouse of digital objects but a seamless integration of documents and tools for research in the intended area.

Keywords: digital libraries, digital archives, Taiwanese history, document processing workflow, digital research environment

[*] Ph.D. Candidate, Dept. of CSIE, National Taiwan University. E-mail: gail@turing.csie.ntu.edu.tw

[**] Professor, Dept. of CSIE, National Taiwan University. E-mail: hsiang@csie.ntu.edu.tw

[***] Post Doctoral Researcher, Dept. of CSIE, National Taiwan University. E-mail: tu@turing.csie.ntu.edu.tw