On Developing Government Official Appointment and Dismissal Databank

Jyi-Shane Liu

University Libraries, Natioanl Chengchi University
64 Sec. 2 Zhih-Nan Rd., Taipei, Taiwan
jsliu@cs.nccu.edu.tw

Abstract. In this paper, we report a databank development project in which structured textual data from historical documents are extracted to provide information access of higher data granularity. The availability of the databank opens up tremendous opportunities for research topics in government personnel systems that were limited by data acquisition difficulty in the past. The project demonstrates the potential of library as data producer in exploiting primary textual resources and developing value-added digital collection.

Keywords: Databank development, Value-added digital collection.

1 Introduction

Over the last twenty years, advances in digital libraries have re-shaped how users expect to use libraries and how libraries respond to changing information needs. While maintaining traditional functions, libraries have become pro-active in creating new values and expanding new services [1][2]. We consider a new role of digital library as data producer that compiles primary research data from textual resources. The notion of databank development has been mostly associated with recording and organizing scientific data [3] and social surveys [4]. As libraries become more active in selecting source materials for digital archiving, the reviewing process presents an opportunity for important subject data to be identified. Many subject data of research values are difficult to access as they are buried in voluminous text of historical documents. Thus, libraries are in a unique position to recognize the opportunity and initiate a data production process to extract subject data from textual resources. By exercising the capability to produce valuable textual data from primary source collection, libraries provide yet another vital support to research communities.

2 Subject Domain

Government gazettes are printed publications available in most major libraries. As bulletins of official announcements, records, codes, and orders, government gazettes provide authoritative government information and are persistent with the existence of governments. Appointment/dismissal orders are authoritative directions issued by the reigning President to appoint/dismiss a named official to/from a named government post. Every instance of appointment/dismissal of a government official is authorized through a written order and is publicly announced.

Given the textual nature of the appointment and dismissal orders, there are a number of ways to access the government personnel changes information. First, paper-based documents can be converted to electronic texts so as to enable full-text search. Words and phrases can be used with Boolean logics to retrieve a subset of flat documents where constraints of linguistic form are satisfied. However, as in common full-text document retrieval, the results require considerable human processing and filtering. Second, electronic texts of appointment/dismissal orders can be further annotated (encoded) by adding semantic information to text pieces with selected tag set. Text encoding enables semantic search/retrieval and allows automatic mapping to entity relation data model. Personnel information as revealed in the appointment/ dismissal orders can be fully specified and accessed. Prerequisite of text encoding includes acquiring convenient annotation tools and training capable annotators.

It is observed that appointment/dismissal orders are compact written forms of personnel changes. Only a very small portion of text pieces does not correspond to entities and relations of personnel changes instances. Text encoding is usually intended to distinguish a small part of informational text pieces from the remaining text. In the case of appointment/dismissal orders, text encoding seems to be over-annotating the entire text. Another problem also arises from the compact nature of the appointment/dismissal orders. Several persons may share the same association of government unit, rank, or title in an order, while only one set of textual pieces are present. These omitted information needs to be inferred and appended so that job change information of each named person is as complete as semantically revealed in the order. Given the purpose of compiling personnel changes information for direct data access and analysis, we take the approach of straightforward extraction and conversion into structured data. In other words, informational textual pieces from the documents are manually identified and keyed into a relational database. Omitted information is also concurrently inferred and the missing textual pieces are supplied during data entry process.

3 Databank Development

The development of government official appointment and dismissal databank has been an on-going project for three years. A full-time librarian was appointed to run the routine work of retrospective data acquisition and inspection, and has been supported by several part-time students as database clerks. The ultimate goal of the project is to complete the retrospective process to include several political regimes related in people or land. Fig. 1 shows the evolution of five political regimes to be included in the databank. Currently, the project has concluded the second stage with complete coverage for the Taipei government (from present back to the year of 1949) on the island of Taiwan and has entered the third stage of retrospective data acquisition for the Nationalist government (from the year of 1949 back to 1925) in mainland China.

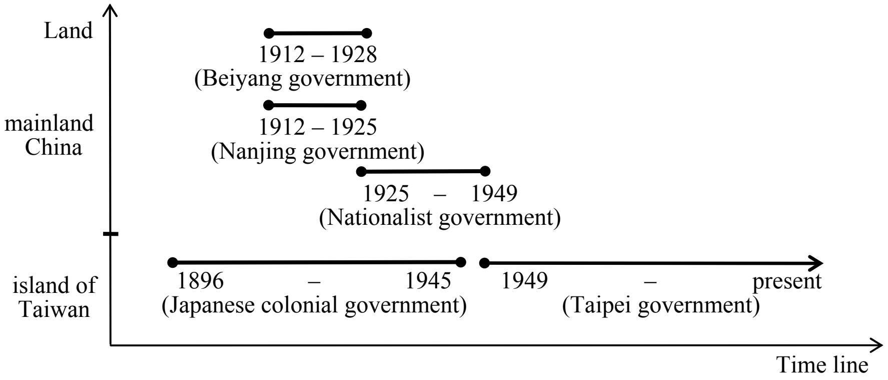
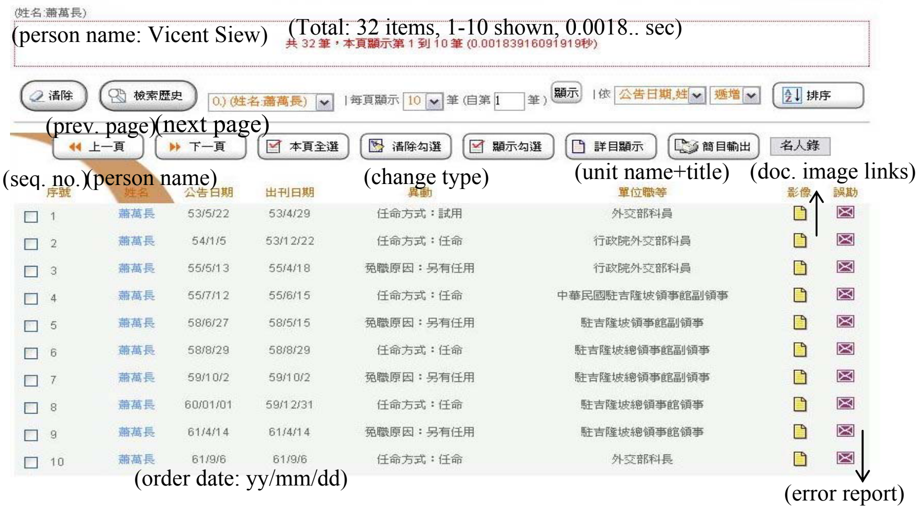


Fig. 1. Evolution of political regimes related in people or land

Three other political regimes are planned to be included in future extension. Nanjing government and Beiyang government both ruled a part of China in the warlord era. Nanjing government took control of the southern China between 1912-25, while the northern China was reigned by Beiyang government between 1912-28. The island of Taiwan was once a Japanese colony between 1896 and 1945. Government gazettes of these political regimes are all available for data acquisition. A further retrospective coverage of the island of Taiwan governance under Qing Dynasty China before 1896 is also considered when relevant documents are identified.



(姓名:蕭萬長)

(person name: Vicent Siew)

(Total: 32 items, 1-10 shown, 0.0018.. sec)

共 32 筆，本頁顯示第 1 到 10 筆 (0.00183916091919秒)

(prev. page)(next page)

(seq. no.)(person name) (change type) (unit name+title) (doc. image links)

序號	姓名	公告日期	出刊日期	異動	單位職等	影像	誤勘
□ 1	蕭萬長	53/5/22	53/4/29	任命方式：試用	外交部科員	📄	✉
□ 2	蕭萬長	54/1/5	53/12/22	任命方式：任命	行政院外交部科員	📄	✉
□ 3	蕭萬長	55/5/13	55/4/18	免職原因：另有任用	行政院外交部科員	📄	✉
□ 4	蕭萬長	55/7/12	55/6/15	任命方式：任命	中華民國駐吉隆坡領事館副領事	📄	✉
□ 5	蕭萬長	58/6/27	58/5/15	免職原因：另有任用	駐吉隆坡領事館副領事	📄	✉
□ 6	蕭萬長	58/8/29	58/8/29	任命方式：任命	駐吉隆坡總領事館副領事	📄	✉
□ 7	蕭萬長	59/10/2	59/10/2	免職原因：另有任用	駐吉隆坡總領事館副領事	📄	✉
□ 8	蕭萬長	60/01/01	59/12/31	任命方式：任命	駐吉隆坡總領事館領事	📄	✉
□ 9	蕭萬長	61/4/14	61/4/14	免職原因：另有任用	駐吉隆坡總領事館領事	📄	✉
□ 10	蕭萬長	61/9/6	61/9/6	任命方式：任命	外交部科長	📄	✉

(order date: yy/mm/dd)

(error report)

Fig. 2. Career path of a named official

The databank currently contains data for official appointment and dismissal dated from 2008 back to 1936. The total number of entity instances is approximately six hundred thousands, which includes two hundred thousands persons, twenty thousands government units, and two thousands job titles. A query interface has been developed for the databank that supports query with Boolean combination of attribute values. As mentioned earlier, a named official's career path can be easily retrieved by a query specifying the official's name as the value of the attribute "person name". Fig. 2 reveals the career path of current Vice President, Vincent Siew, who started his government work as a staff of the ministry of foreign affairs in 1964. Each appointment/dismissal instance is also linked to the document image where the order appeared and provides a chance for users to detect and report errors.

An essential implication of recording personnel change information in a structured collection of data sets is the ability to provide higher data granularity for more effective information use. In other words, the databank provides fine-grained information that can be integrated and analyzed as needed to reveal unknown aspects and trends of subject matters. Such a data analysis utility is not possible with document retrieval.

4 Conclusion

This paper reports the development of government official appointment and dismissal databank. The databank provides higher data granularity of government personnel information and enables dynamic data exploration that helps discover facts and knowledge with multi-faceted analytical investigation. The databank development work provides an example of how digital collection of document images can be extended to create new values for users with more effective information use. In particular, structured textual data of great value to special subject domains can be extracted from historical documents and other primary source collections to facilitate creative research in humanities and social sciences. With direct access to the collected primary source materials and the expertise of information organization, libraries are in a unique position to play the role of data producer and develop textual databanks that help reveal information and knowledge hidden within volumes of documents. The proposed direction not only upholds libraries' core value of information service but also creates a valuable niche for libraries in the shifting information landscape.

References

1. Brogan, M.: A Survey of Digital Library Aggregation Services. Technical report, The Digital Library Federation (2003)
2. ApSimon, J.,and NDAC Working Group: Building Infrastructure for Access to and Preservation of Research Data. Technical report, Social Sciences and Humanities Research Council of Canada (2002)
3. Benson, D.A., Karsch-Mizrachi, I., et al.: GenBank. Nucleic Acids Research 28(1), 15–18 (2000)
4. Kavaliunas, J.: Census 2000 Data Products. Government Information Quarterly 17(2), 209–222 (2000)