Sharing Historical Records on the Internet
Experiments at the Japan Center for Asian Historical Records

Shohei MUTA
Japan Center for Asian Historical Records
National Archives of Japan
smuta@archives.go.jp

Abstract:
The Japan Center for Asian Historical Records of the National Archives of Japan (JACAR) was established in November 2001 for the purpose of providing people of all over the world with access through the Internet to official records kept by various ministries and agencies of the Japanese government on modern history of relations between Japan and various countries, primarily neighboring Asian countries. In order to fulfill its mission, the center introduced very unique technology and systems such as the DjVu document compression technology; a catalog system relying on the General International Standard Archival Description (ISAD (G)) and Dublin Core; and a digital dictionary for synonymous, related words, and English translations specialized for Japanese modern history.

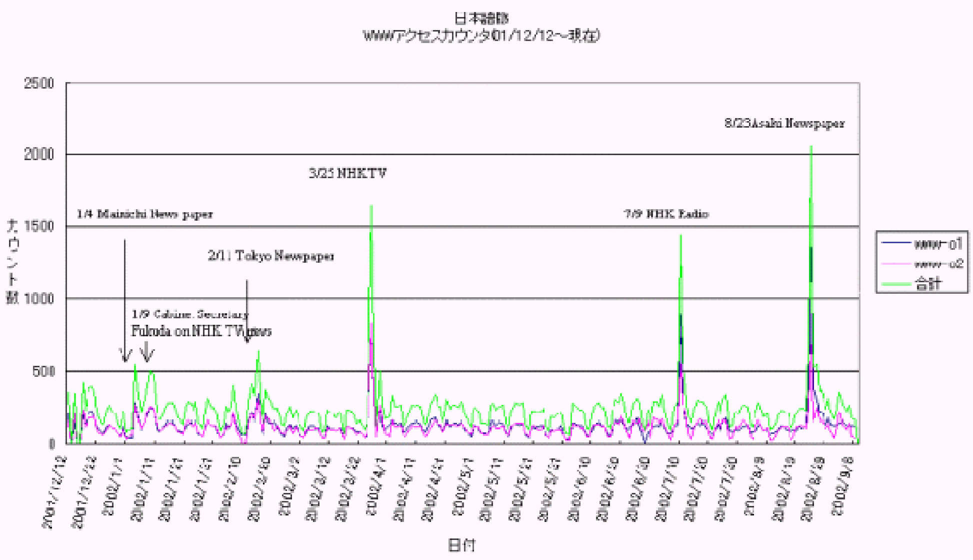
1. Introduction

JACAR (http://www.jacar.go.jp) was established on 30th of November 2001 as a part of an organ of the National Archives of Japan. As of August 2002, the center provides about 2.3 million pieces of image data and a catalog database of 160,000 items, which will be augmented on an ongoing basis. Just within two years after the Cabinet decision to establish the center on November 1999, the center was urged to come up with an information system, which is able to handle huge image and catalogue data, and a unified cataloging system, and working procedures with archival institutions participating in the project. Finally, the information system, which will be introduced later part, was adopted.

2. Access status

2. 1. Access statistics

There are about 80,000 accesses in 10 months. As the following chart shows that exception of occasional steep increases after media coverage, access rarely fluctuated. Other interesting characteristic is that there are constant decreases of access on every Wednesdays. In Japan, most academic institutions hold faculty meetings on Wednesdays. Hearings and a questionnaire survey also support our observation that the users are mostly academics accessing from their university offices. Our new challenge is to increase access by general users.



JACAR Access Count

2. 2. Search terms

The center keeps logs only for the purpose of improving our information service. During the first six months, about 60,000 terms are used for keyword search. After eliminating duplications, about 18,000 historical terms are extracted and examined. The followings are the first ten terms on the list.

1. 朝鮮 or Korea 747 times, 2. 慰安婦 or comfort women 523 times, 3.台湾 or Taiwan 450 times, 4. 徴兵検査 or conscription examination 437 times, 5. 太平洋戦争 Pacific War 389 times, 6. 南京 or Nanjing 389 times, 7. 上海 or Shanghai 340 times, 8. オーストラリア or Australia 336 times, 9. 中国 or China 331 times, 10. 満州 or Manchuria 303 times

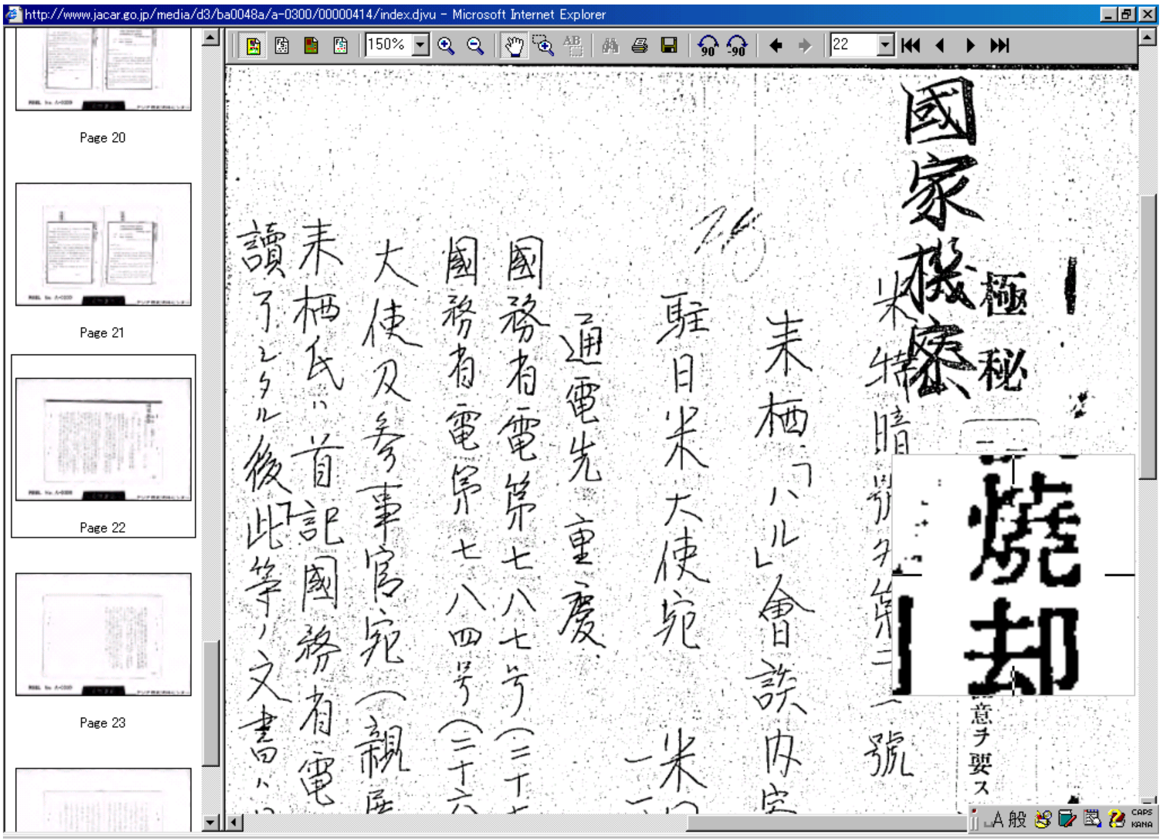
There are very few unexpected terms used and most terms seem to be used for academic research. These terms are later incorporated into our digital dictionary as base terms.

2. 3. Materials accessible through the Internet

Following the Cabinet decision of November 1999, the center provides official records from collections kept by three major archival institutions, the National Archives of Japan, the Diplomatic Office of the Ministry of

Foreign Affair, and the Library of the National Institute of Defense Studies, which keeps former Army and Navy records. The materials are official documents and other records related to Japan's past relationship with various neighboring Asian countries and regions, spanning from the begging of the Meiji era to the end of the Pacific War. The total data of Asia-related records in three institutions are estimated about 27 million image data and 1.2 million catalog data.

DjVu image sample（Diplomatic Record Office : Reference No.B20020307503）


http://www.jacar.go.jp/media/d3/ba0048a/a-0300/00000414/index.djvu - Microsoft Internet Explorer

國家機密 極秘

焼却

意ヲ要ス

米特暗號第二號

來栖「ハル」會談内容

駐日米大使宛（米

通電先 重慶

國務省電第七八七号（三十

國務省電第七八四号（二十六

大使及参事官宛（親展

來栖氏ハ首記國務省電

讀了シタル後此等ノ文書ハ

The above image sample shows that there are special notes such as「國家機密」(State Secret), 「極秘」(Top Secret), and「用済後焼却スベシ」(Incinerate after reading). These documents are not originally intended for public disclosure and to be considered as valuable historical documents. This sample document, categorized as "Special Information"「特殊情報」by the Diplomatic Record Office, is a translation of a deciphered diplomatic correspondence sent by Secretary Hull to Ambassador Grew in Tokyo just after the handing of the so called "Hull Note" at the U.S. State Department to Japanese Ambassadors Nomura and Kurusu in November 1941.

3. Outline of the JACAR information system

3.1. Function

Through the Internet and utilizing the advanced technology, the function of the information system aims to provide anyone with the means to search, print, and download image data of records at anytime, from anywhere, at no charge. To realize the above function, the following points are considered:

Employment of user oriented search system

Securing impartiality and objectivity of search and cataloging system

To provide multi-lingual search system, at least in English

Employment of synonymous, related words, and English dictionary to assist search not only for general public but also for foreign users

Use of high document compression technology to overcome still inadequate information infrastructure in Japan and abroad

In the following section, two major information systems that are critical for the success of JACAR's information system are described; the search system with digital dictionary and the DjVu image compression technology.

3.2. Conceptual framework of the JACAR search system

Free keywords search and full text search are standard methods in the Internet. However, historical terms commonly used in Japanese modern history are not always the same as the terms used in historical records. For example, **太平洋戦争** "*Taiheiyo Senso*" (Pacific War) does not appear in official documents because the term was used after the War. The official naming of the War by the Japanese government was **大東亜戦争** "*Daitoa Senso*" or the Great East Asian War. In order to fill these gaps, the JACAR historical dictionary, consists of synonymous, related-words, and English translations of key terms, was developed.

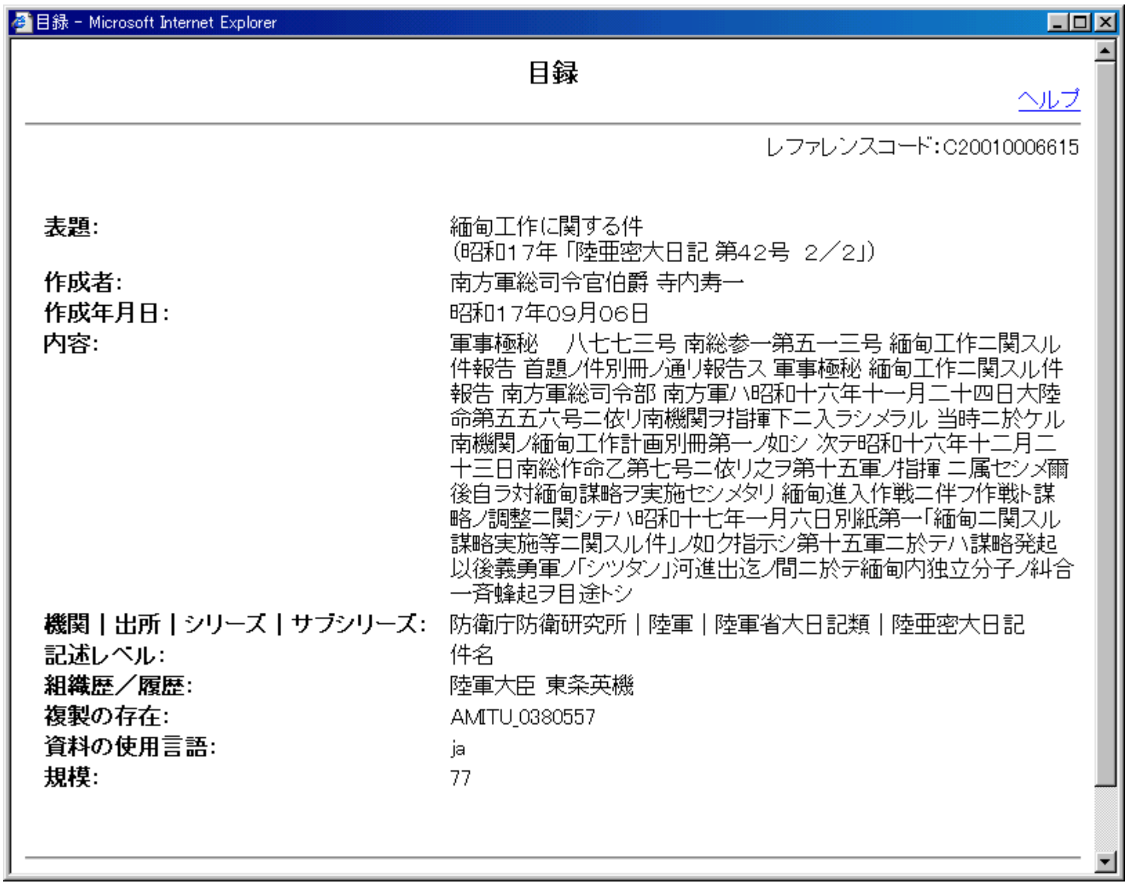
Even with a well-compiled dictionary, without relevant information in catalogs matching with the searched terms, documents will not be retrieved. Titles of the most documents do not always have such key terms like "*Daitoa Senso*." They are usually found inside the text body. Therefore, it is desirable to have all the texts to be converted into machine-readable formats. However, the task of digitalizing Japanese texts is impracticable considering the amount of texts and costs for digitalization without relying on OCR, which is not reliable for reading Japanese handwritten documents mixed in kana (Japanese phonetic syllabaries) and kanji (Japanese characters) format. In order to overcome this constraint, the element of catalog, 'Description' consists of approximately the first 300 characters of each body texts, thus not only the title but also a portion of the text becomes subject of search.

3. 2. 1. Catalog elements

In order to determine the structure of the catalog elements, the Dublin Core and the ISAD (G) (General International Standard Archival Description) are studied. ISAD (G) is used as the basis for organizing independently structured cataloging system of each archival institution into the same hierarchical structure. Thus users are able to search collections of each institution following its hierarchical order as well as to conduct cross-sectional search through the same hierarchically leveled catalogs.

The catalog elements are important guide for searching through historical records. Especially, following five elements are basic ‘access points’ for historical documents; “topics,” “dates,” “places,” “persons,” and “organizations.” ISAD (G) emphases in importance of “access points” for retrieval of documents and chooses three basic “access points,” the “corporate bodies,” “persons,” and “families.” Information related to the five access points are extracted from documents and inputted into the main elements of JACAR catalog, topics are in “Titles,” persons are in “Creator,” dates are in ‘Date,’ topics, dates, places, persons, organizations are in “Description,” and organizations are in “Organization.” English search system uses English catalogs translated from Japanese catalogs exception of “Description” element, which is costly to translate as a whole.

Catalog data sample



目録

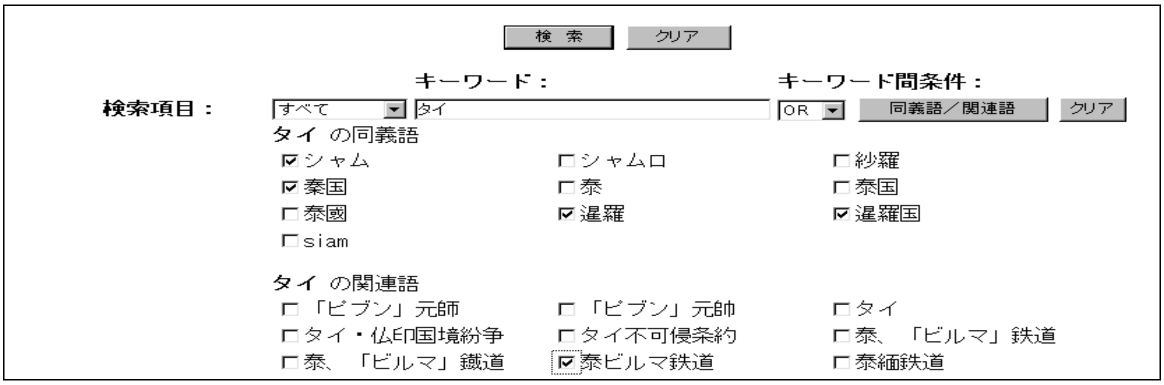
レファレンスコード:C20010006615

表題:	緬甸工作に関する件 (昭和17年「陸亜密大日記 第42号 2／2」)
作成者:	南方軍総司令官伯爵 寺内寿一
作成年月日:	昭和17年09月06日
内容:	軍事極秘　八七七三号 南総参一第五一三号 緬甸工作ニ関スル件報告 首題ノ件別冊ノ通リ報告ス 軍事極秘 緬甸工作ニ関スル件報告 南方軍総司令部 南方軍ハ昭和十六年十一月二十四日大陸命第五五六号ニ依リ南機関ヲ指揮下ニ入ラシメラル 当時ニ於ケル南機関ノ緬甸工作計画別冊第一ノ如シ 次テ昭和十六年十二月二十三日南総作命乙第七号ニ依リ之ヲ第十五軍ノ指揮 ニ属セシメ爾後自ラ対緬甸謀略ヲ実施セシメタリ 緬甸進入作戦ニ伴フ作戦ト謀略ノ調整ニ関シテハ昭和十七年一月六日別紙第一「緬甸ニ関スル謀略実施等ニ関スル件」ノ如ク指示シ第十五軍ニ於テハ謀略発起以後義勇軍ノ「シツタン」河進出迄ノ間ニ於テ緬甸内独立分子ノ糾合一斉蜂起ヲ目途トシ
機関 | 出所 | シリーズ | サブシリーズ:	防衛庁防衛研究所 | 陸軍 | 陸軍省大日記類 | 陸亜密大日記
記述レベル:	件名
組織歴／履歴:	陸軍大臣 東条英機
複製の存在:	AMITU_0380557
資料の使用言語:	ja
規模:	77

3. 2. 2. Synonymous, related words, and English dictionary

The most important point of creating the JACAR historical term dictionary is to secure its objectivity and impartiality of selection of terms. At first, the terms, which are going to be the keywords for search, are selected from standard historical dictionaries and other reference books. Since high school students are included in our targeted users, a historical terms dictionary for high school students is also included as a reference source. In addition, morphological analysis was conducted for 100,000 titles and the results were incorporated into the dictionary. As of September 2002, there are 5,600 keywords. English keywords are selected from standard Japanese history books in English such as *The Cambridge History of Japan* and Japanese-English dictionaries for former army and navy published before the Pacific War. Postgraduate students majoring in modern Japanese history, military affairs, political science, and Asian history do all the editing and compiling works. A committee of professors in modern history and military specialists reviews the outcome of their works. Keyword search activating the dictionary functions as follows.

Sample of the dictionary contents viewed in the Web page



検　索　クリア

検索項目：　すべて ▼　キーワード：タイ　キーワード間条件：OR ▼　同義語／関連語　クリア

タイ の同義語

☑シャム　□シャムロ　□紗羅
☑秦国　□泰　□泰国
□泰國　☑暹羅　☑暹羅国
□siam

タイ の関連語

□「ピブン」元師　□「ピブン」元帥　□タイ
□タイ・仏印国境紛争　□タイ不可侵条約　□泰、「ビルマ」鉄道
□泰、「ビルマ」鐵道　☑泰ビルマ鉄道　□泰緬鉄道

Input「タイ」(Thailand) as a keyword and activate the synonymous/related word dictionary by clicking「同義語／関連語」button. The list of synonymous appears as the column below 「タイの同義語」or synonymous of Thailand such as「シャム」「暹羅」「泰国」「泰國」「siam」. In the column below「タイの関連語」or related-words of Thailand, following terms are listed「ピブン元帥」or Marshal Phibun of Thailand、「泰緬鉄道」or Thailand-Burma railway and so on. Users can choose any term just checking buttons. In the English search page, input “Thailand” and activate “Synonymous” or/and “Related-Words” dictionaries by checking buttons, the search will be expanded to the Japanese catalog data as well as English catalog data.

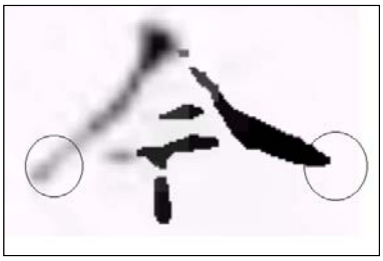
3.3. Image data processing

Each archives holding original documents creates microfilms first and converts them to digital format. Microfilms are used for backup by original holding institutions. After examining costs of direct conversion from original documents to digital formats, it turned out to be more economical, efficient, and less manpower required to microfilm the documents and then to digitalize by automated microfilm-scanning machines.

The test results for conversion show that a bi-level 400dpi sample image occupies about 1.5MG in average in JPEG format. Since total estimated volume of the data were 32 million image data. It would require 48TB of storage memory and an additional backup storage. Considering the size of available funds, it was unfeasible. It was also obvious from the online access test that huge image data transfer is impracticable even using image compression technologies such as JPEG and PDF formats. After extensive research and hearings, the DjVu format was adopted. A group of historians and image data specialists worked out the specifics for digital conversion, which would provide enough resolution and gradation for historical research. Since most documents are in bi-level image and as far as users can read document texts, it serves primary objective of the system. It was then decided that digitalize specifications from microfilms are bi-level, 400 dpi, and in TIFF format, which is an almost standard format. TIFF files are later converted to DjVu format, which has higher compression ratio for bi-level data and provides useful functions such as zooming and panning, magnifying, browsing of pages and so on by a free plug-in for commonly used web browsers such as Netscape and Internet Explorer.

DjVu format has 10 to 20 times better compression ratio for bi-level text data than JPEG format. It separates images into two parts; foreground texts and background images, and compresses with two different methods. Texts and drawings (foreground) are compressed in a lossless scheme called JB2 and background images are compressed in a lossy scheme called IW44.

Gray scale image sample



This is the image of character「今」,1,200% magnified and adjusted to the present size. Inside the left side circle shows lossy compression results and inside the right side circle shows losseless compression result. Because of the character 「今」was written in brushstrokes and converted into gray scale image, there is confusion in foreground and background separation. The sample clearly shows the difference between two types of compression schemes. However, it is hard to see the difference in the original size. Nonetheless, this problem does not occur with bi-level image, which is JACAR's standard format.

Though there were opinions against adoption of this new and not yet tested DjVu format, it was irresistible fact that DjVu is the most suitable compression technology for JACAR's mission. However, considering DjVu's uncertainty, we provide JPEG format for users who cannot use DjVu viewer and keep TIFF files in order to meet new developments in image compression technology. As of August 2002, 2.3 million images occupy about 400 GB. Average file size is about 175 KB. Because of an adoption of bi-level image scheme, background date size is very small, which enable us to achieve very high compression ratio than we had estimated.

4. Conclusion

The JACAR information system can be characterized in two; user oriented search methods such as layered search, keyword search, and expert search using the JACAR dictionary, and English search systems; and the data transmission scheme using advanced image compression technology. However, the system is not aimed at transmitting three-dimensional object images commonly used in a "digital museum." The JACAR image based digital format and catalog system modeled after ISAD (G) and Dublin Core is more suitable for digital archives and libraries, where most of their materials are in the form of bi-level texts and are difficult to digitalize their contents because of handwritings and variety of printing styles.

Reference

ISAD (G): general international standard archival description second edition adopted by the Committee on Descriptive Standards Stockholm, Sweden, 19-22 September 1999.
www.ica.org/eng/mb/cds/descriptivestandards.html

Dublin Core nitsuite, Sugimoto Shigeo,"Joho Kanri" (Vol.45 No.4, July2002)

Using Dublin Core, Diane Hillmann,
http://dublincore.org/documents/2001/04/12/usageguide/

Compression of bi-level images, Adolf Knoll,
http://www.nkp.cz/start/knihcin/digit/vav/bi-level/Compression_Bi-level_images.html

For DjVu compression technology please refer to http://lizardtech.com