NII Portal Sites for the Digital Silk Roads Project

Asanobu Kitamoto, Eric Platon, Frederic Andres and Takeo Yamamoto

National Institute of Informatics

kitamoto@nii.ac.jp, platon@nii.ac.jp, andres@nii.ac.jp, ty@nii.ac.jp

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN

Abstract

This paper briefly describes the basic concepts and current status of portal sites that National Institute of Informatics (NII) is involved in the design and operation for the Digital Silk Roads (DSR) Project. These portal sites aim at disseminating and sharing cultural heritage information related to Silk Roads. This paper introduces two portal sites, namely Toyo Bunko Portal, and Advanced Scientific Portal for International Cooperation on Digital Silk Roads (ASPICO-DSR). This paper also addresses other related sites and programs, namely Cultural Heritage Online and ECAI and concludes with a future plan.

Keywords

Portal Site, Toyo Bunko Image and Manuscript Database, Advanced Scientific Portal for International Cooperation on Digital Silk Roads

1. Introduction

National Institute of Informatics (NII) made an agreement with UNESCO on the Digital Silk Roads Initiative Framework (DSRIF) in 2001, and following the agreement, we started an international collaborative research project called the Digital Silk Roads. The purpose of the DSR project is to propose a new approach for the preservation and presentation of the huge amount of cultural heritage by means of cooperative work between informatics communities and cultural studies communities. The role of informatics communities is to provide efficient information infrastructure to assist cultural studies on the Silk Roads. The project members consist of researchers from more than ten countries including many central Asian countries. Under this framework, several research programs have already started, and more information about these projects is available on the portal site of the DSR project[1].

Other portal sites are also under development and this paper introduces two of them, namely (1) Toyo Bunko Portal and (2) Advanced Scientific Portal for International Cooperation on Digital Silk Roads (ASPICO-DSR). These portal sites are integral parts of the DSR project to achieve the goal of providing global information infrastructure for disseminating and sharing cultural heritage resources. Those portal sites respect the basic policy of free and efficient access to digital cultural resources, but they pursue different targets with their own designs. The former project focuses more on the analysis and annotation of digital cultural resources, while the latter focuses more on collaborative working environment and knowledge management.

In the following, we first introduce Toyo Bunko Portal in Section 2 and ASPICO-DSR in Section 3. Then we address other related sites and programs in Section 4, and conclude the paper in Section 5.

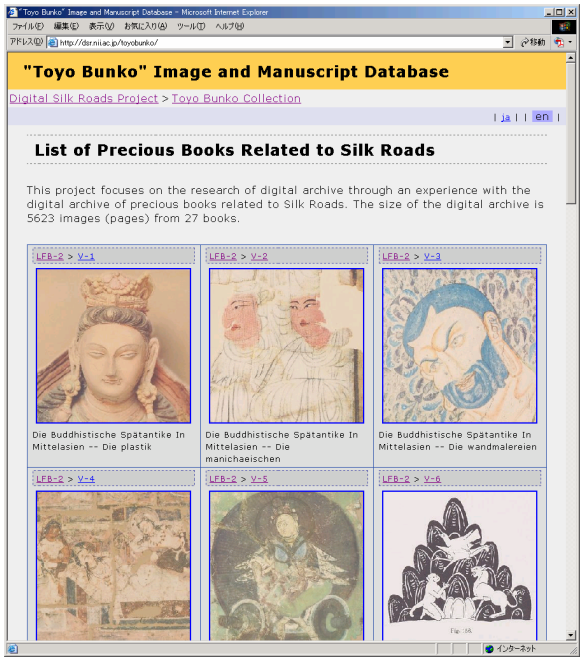

Fig. 1 Toyo Bunko Portal Site.

2. Toyo Bunko Portal[2]

[1] http://dsr.nii.ac.jp/

[2] http://dsr.nii.ac.jp/toyobunko/

2.1 Overview

The main content of the Toyo Bunko Portal is Toyo Bunko Image and Manuscript Database (Fig. 1), which is a collaborative project with the Toyo Bunko (Oriental Library), a leading library in the field of Asian studies. Its collection amounts to 880,000 books of historical importance, but unfortunately, some of the books are "invisible" from the general public because of limited accessibility to those books mainly for preservation and safety purposes. To improve accessibility to those books, we suggest that the digital archive is the best solution in terms of both preservation and accessibility.

Among its collection of historical importance, an especially interesting collection is "Morrison Library," which consists of 24,000 books about China and Asia written in several European languages. Regarding its relevance, scale and coverage, we decided to start our digital archive project from the Morrison Library, and initiated the digitization of rare books in FY 2002. In two years we digitized 5,723 pages from 27 rare books, ranging from the report of academic explorations to more personal travelogues. Those books are digitized using professional digital cameras with the resolution of either 4,072 x 4,072 pixels or 10,500 x 12,600 pixels. On this digital archive we are planning to purse two research directions.

The first direction concerns the application of automatic analysis processes such as optical character recognition (OCR), machine translation and image processing on digitized documents. Our motivation behind this direction is the need for the management of large number of books. To increase the number of books in the digital archive, we need to put more emphasis on speed than precision, and our idea is to increase the precision by means of post-processing of the result of automatic analysis, such as the correction of technical terms using a specially-designed tool for proofreading. The second direction is the collaborative annotation environment for digital cultural resources. We begin with closed annotation by domain experts, but in the future we plan to establish a mechanism for soliciting collective annotation in a collaborative environment.

2.2 Basic Design

The design goal of the Toyo Bunko Portal is to improve accessibility to the contents of rare books. Hence the web site should provide both navigational links for effective browsing and various searching mechanisms for jumping directly to the information in need. As navigational links we provide four types:

1. Language Links
   Toyo Bunko Portal is designed from the beginning to allow multilingual access to the contents of books. For this purpose we place navigational links on every page to switch the language on display. The choice of language is also performed on the Web server using content negotiation and the tracking of preferred language of the user.

2. Page Links
   Page links are probably the most useful tool as the navigation of books. For example, reading a book from the front cover to the back cover is just about following the "next page" link on every page to reach the final page. We also provide some links for jumping into distant pages for quick browsing over the book.

3. Structure Links
   In digital archives we do not have to stick to the physical construct of a book, like a page, but have more freedom to focus on the logical chunk of information such as a paragraph, a section, and a chapter. Links to these logical constructs of a book, namely structure links, are useful with appropriate labels. At this moment these links are not available because of the requirement of manual annotation, but they will be available in the near future.

4. Text Links
   Text links are used for navigating through a book based on relevant words. This is similar to referring to the index of a book and jumping to related pages that contain the word. This type of links is still under development, but we instead provide a similar functionality by means of a full text search engine called Namazu. [1].

We also provide other searching mechanisms like a map-based search. This is a simple clickable map and on clicking the place you are redirected to a full text search engine to find relevant pages containing the name of the place. Fig. 2 is the interface of map-based search, and on clicking the place of Kashgar, you can find 73 pages that contain the word Kashgar. More elaborate searching mechanisms, such as those combined with Geographic Information Systems (GIS). are required in the future to help cultural studies based on regions, countries and roads.

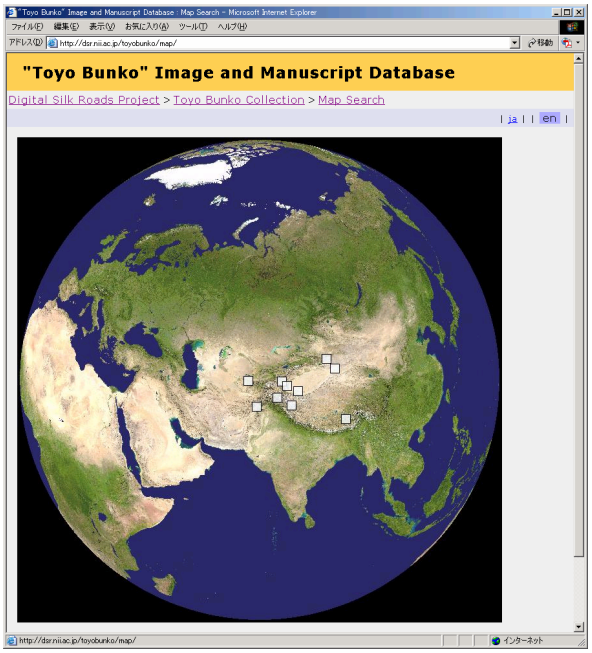

Fig. 2 The map-based search interface.

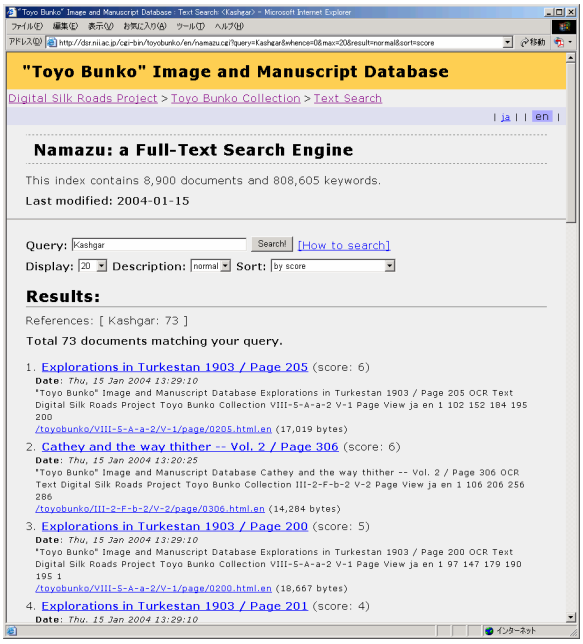

Fig. 3 The result of map-based search.

2.3 Text Extraction

Text extraction and processing is the most important part of this portal site, since the extraction of contents as an electronic text makes it possible to search into the book based on keywords. This facilitates the reuse of information in the book, even when the index of the book is not available.

Obviously the best way for creating an electronic text is the manual input of the full text. Examples of the manual input of an electronic text include Project Gutenberg [2] which focuses on light literature, heavy literature and references, and Aozora Bunko [3] which focuses on Japanese literature. They have many contributors that take the task of text input and proofreading. This kind of collaborative model works for the masterpieces of literature on which you can expect many readers. Other publications, such as Japanese classic literature of academic importance, sometimes have the database of electronic full text proofread by domain experts [4]. In terms of academic reports and minor publications, however, the manual input of text is too costly and hardly feasible.

Hence we applied optical character recognition (OCR) to extract textual information from the book [5]. This text information is displayed on the page view so that a user can compare erroneous OCR text with the real digital image in order to check the correctness of the full text. The full text can be further utilized for text processing such as full text search and multilingual dictionaries.

2.4 The Future of Toyo Bunko Portal

This portal site has almost completed its initial development and is ready for public access. The opening of the portal site, however, is delayed due to non-technical issues, which we hope to solve in the near future.

The second stage of development includes the collective annotation of cultural heritage information through the web browser, and more advanced text processing on the extracted text. The to-do list also includes the application of image processing for the automatic extraction of illustrations, paintings and photographs from pages to create the index or the search engine of graphical elements in a book.

We still have a lot of cultural heritage in the library but they are just sleeping in the dark room of libraries and waiting for accesses from everyone. The public access may be difficult on physical books due to preservation reasons but the access is easier on digitized books.

Many researchers also point out that the careful inspection of digitized images often result in new academic discovery which have not been possible on real cultural heritage. This is because a study with real cultural heritage is often restricted by various real-world conditions such as distance, lighting, time, and memory. With digitized images, those restrictions can be removed, and this fact leads to a new line of cultural studies in which the data acquisition process and the data analysis process is separated by means of information technology such as image processing and database technology. Hence we believe that the

development of digital archives is a fundamentally important activity to enhance our knowledge through the discovery of buried knowledge in books and cultural heritage in general.

3. ASPICO-DSR portal[3]

In the Digital Silk Roads Initiative, endeavors conducted by the UNESCO and the National Institute of Informatics will breed multimedia and multilingual contents about these millenarian paths. This cross-disciplinary collaboration between international researchers and experts needs a common access point to share and build together quality knowledge. Hence, a web-based solution was proposed, namely the Advanced Scientific Portal for International Cooperation on Digital Silk Roads (ASPICO-DSR), based on the open project 'Chef' [6]. The following parts aim at presenting the main features of the project and its future evolution.

3.1 Expert Membership and Public Area

The ASPICO-DSR portal addresses two points of view for its cultural contents. First, the Member Area provides researchers and experts with a work environment (named Work Space), in order to create contents and offer their expertise in interpretation and comments. On the other hand, the Public Area lets anyone world-wide accessing available cultural contents from member works, through a category-based search engine that exploits comments to index resources.

The remainder of this part about ASPICO-DSR will describe the Member Area features. Although integrally part of the framework, the Public Area can be thought of as a special search engine about Silk Roads. In this regards, it is similar to general tools such as Google™ or AltaVista™. Currently, the main difference consists in user orientation to choose categories and refine them steadily to the expected result, rather than full text search mode.

3.2 Content Creation Facilities and Committee Quality Validation

The ASPICO-DSR framework embodies first services to manage, share and comment multimedia resources. On their Work Space (c.f. 3.1), members can send their data to the portal for temporary or permanent storage, and optionally allow any other member to access this data by defining individual or group authorization. When data is available on the portal Work Space, members can then create description, interpretation or any personal comment.

When comments are satisfactory, members apply to the ASPICO-DSR committee for publication of their resource candidates on the Public Area. This review conducted by elected members will ensure the overall information quality and public access to concrete results.

So as implementing this review process, a data versioning system was designed to clearly distinguish the three identified states as follow.

- Temporary state (Red metadata): Raw data is on the portal and needs comments from expert members for further exploitation (discussion and publication). Such data is only accessible from the owner member Work Space.
- Clean and Commented state (Orange metadata): The resource consist now in a set of data and its corresponding comments. The resource is ready for review by the committee before publication. It is still only in the member Work Space.
- Validated state (Green metadata): The resource was reviewed by the ASPICO-DSR quality committee and is now available on the Public Area.

By this simple color code, members can follow the current state of their resources and work accordingly with the community.

3.3 Communication and Collaboration Facilities

So as to ease and foster wider collaboration, the second feature of the ASPICO-DSR portal gathers communication tools. Currently, the solution inherited from Chef supports:

- An internal version of instant messaging to discuss synchronously.
- A forum to thread area of interest discussions about resources or various subjects.
- A file exchange area to conveniently transfer data among members.

These tools should let members work together freely. Farther, members can also create and control specific web-pages for collaboration with fellows. They can decide which communication tools and resources can be exploited on this shared area (named Work Site). This capability is to be extended to public area access, in order to allow members showing their personal working data.

[3] http://www.aspico-dsr.org/

3.4 Future Advanced Tools

The last features that should be available steadily are original tools to search content-based data and their corresponding services to assist members in such indexation process. In fact, the current solution allows searching resources based on the category they belong to. The content-based future search provides a finer search grain by linking objects included in resources with related categories. For instance, the following image (from the future collection of the portal) can be categorized as in Paintings/History/Asia.

The selection of some or all categories in the search engine would at least display this picture, but this taxonomy binds usage to global requests about the resource. The content-based framework aims at adding relevant categories such as 'Horse' to link directly to the horse contained into the picture. More abstract categories could also be defined such as 'Fight Scene'. This last feature is frequently required to describe representations of Buddha (sitting, standing, etc). Future works at NII will address a way to correctly link objects and comments in these resources.

4. Related Sites and Projects

4.1 Cultural Heritage Online

Cultural Heritage Online is the project headed by the Agency for Cultural Affairs and Ministry of Public Management, Home Affairs, Posts and Telecommunications with the purpose of building a portal site for more than 1,000 museums in Japan. This site provides basic information about museums and their selected collection of cultural artifacts. Basic information is provided with both an annotation such as the creator, time of creation, and expert's description, and a low resolution image which takes the artifacts possibly from multiple directions to illustrate the appearance of the artifact. Although this set of information is not enough for scientific study, this site will help for the general public to access information about cultural resources that are not easily searchable and viewable across museums at this moment.

For improving accessibility, interactive searching and browsing of the database is the key technological challenge. This portal solves this issue through the introduction of a powerful engine called GETA [4], which has been developed in National Institute of Informatics and other organizations with support from IPA (information technology promotion agency). This engine provides what they call associative search, which means that users can browse through the database following the association of data calculated from the appearance and distribution of words.

NII has a close relationship with Cultural Heritage Online in terms of the support of infrastructure and the submission of data (Toyo Bunko resources). This portal site may also serve as a gateway to the Digital Silk Roads programs.

4.2. ECAI

The Electronic Cultural Atlas Initiative (ECAI) [7] implements the historical atlas paradigm to find and access published world-wide cultural contents. It relies on a Geographical Information System, extended with the time dimension.

Projects entering the ECAI community just need formatting their resources metadata to the group norm (conversion tools are provided), so that all are seamlessly searchable and accessible through the portal. When visitors query the ECAI engine and fields of interest are found, an interactive map displays relevant information along the time axis. Selection of peculiar places or dates modifies the context to show corresponding data.

ECAI and ASPICO-DSR share the common mission of cultural heritage availability, and it leads to close collaboration in the future. On the one hand, ASPICO-DSR should share its contents and offer its complementary dimensions in domains of interest (Architecture, Sociology, etc) about the Silk Roads. On the other hand, ECAI should provide its contents and interactive map tools. These tools will allow the starting ASPICO-DSR community to encompass space information and link resources directly with their original or discovery time and place.

5. Conclusion

[4] http://geta.ex.nii.ac.jp/

For the time being, the issue of system integration of two portal sites introduced above has a low priority, since we are still exploring new models on the future of digital archives especially for cultural heritage. We are still in a premature stage, and too rigid design of the system in the initial stage may curb the flourish of innovative ideas in the future. Hence we will pursue various approaches at this moment, and integrate the portal sites in the future into the best model we will find. Especially premature area of research and development is the realization of collaborative working (annotation and validation) environment, which is one of our main concerns. In principle, we rely heavily on open source software, open protocols, and open formats to keep the independence of our project from proprietary movements, and will start system integration gradually from individual elements to the whole.

Acknowledgment

Asanobu Kitamoto and Takeo Yamamoto thanks Dr. Yoshinobu Shiba and Dr. Issei Tanaka in the Toyo Bunko for their helpful support to the digitization of precious books in the Toyo Bunko.

References

[1] Namazu Project, http://www.namazu.org/.
[2] Project Gutenberg, http://promo.net/pg/.
[3] Aozora Bunko, http://www.aozora.gr.jp/.
[4] Digital Library of Japanese Classic Literature, National Institute of Japanese Literature, http://www.nijl.ac.jp/contents/d_library/.
[5] Sonoko Sato, Asanobu Kitamoto, Yoshinobu Shiba, Issei Tanaka, Kinji Ono, and Takeo Yamamoto, “Multilingual DSR Document Archive: Digitization, Automatic Multilingual Indexing and Collaborative Thesaurus Construction”, Proc. of Nara Symposium for Digital Silk Roads, This volume, 2003.
[6] Hardin, J. et al.: ‘The CompreHensive CollaborativE Framework’ (CHEF), http://www.chefproject.org, University of Michigan, accessed on December 2003.
[7] Lancaster, L et al.: The ‘Electronic Cultural Atlas Initiative’ (ECAI), http://ecai.org, ECAI with support from Berkeley California University's International and Area Studies, accessed on December 2003.