Digitization of Rare Books and Documents

KITAMOTO Asanobu

National Institute of Informatics 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 1018430, JAPAN kitamoto@nii.ac.jp

July 14th- 24th 2004 JSPS Asian Science Seminar 2004, NII, Tokyo

Two Reasons for Taking This Lecture

You will be (or are) involved in digital archive projects. You will be a user of digital archives.

#In any case, you need to know about digital archives.

Three Reasons for the Importance of Rare Books

#Basis of research and education. #The symbol of culture. #Compilation of knowledge which human beings ever acquired.

Four Reasons for the Digital Archive of Rare Books

#Preservation #Accessibility #Searchability #Evolution

#Because they are in danger!





#Preservation is important.

#Because we can't see them!

Books are kept on bookshelves of the room

¥ You cannot read or even see them!

#Accessibility is important.

#Because we can't search them!

The title and description does not help. We are used to Googlestyle full text search engine.

#Searchability is important.

#Because we want to share!

The developer gives additional information. Other people give contribution for their knowledge.

#Evolution is important.

Outline

- 1. Introduction
- 2. Input Devices
- 3. Image Processing and Color Management
- 4. Metadata and Annotation
- 5. Database Systems
- 6. World Wide Web and Case Studies
- 7. Future Direction

What are Rare Books?

#Books that contain academically valuable contents. #Long time has passed since publication. #Only a limited number of books survived over time.

Motivation

Preserve the books as cultural heritage for future generations. # Annotate books to create a rich collection of knowledge about books. # Search and browse books for the contents of text and image. # Let everyone read valuable books anywhere and anytime.

Issues

- How can we ensure preservation? ... Part 2 and Part 3.
- How can we evolve knowledge? ...
 Part 4.
 - 3. How can we enhance searchability? ... Part 5.
 - 4. How can we improve accessibility? Part 6.

Introduction

#Four important issues of digital archive : preservation, evolution, searchability, and accessibility. #Public digital archives are sensitive to the issue of copyright, or intellectual property rights (IPR).

Preservation

A major role of the digital archive is the preservation of cultural heritage. # When a real book is lost or damaged, a digital data can be used for the reconstruction of the original. # Quality control is required throughout the digitization workflow to keep the fidelity of data.

Evolution

Digital archive is not just about digitizing real cultural resources. # Digital archive evolves by annotating related information to resources. # Cultural resources are multi-facet, hence annotation requires collaborative work of many experts.

Searchability

Electronic texts, or Etexts, offers high searchability in terms of the contents of books.

Image processing offers high searchability into image contents.
Extracted information is fed into database systems to improve and speed-up searching functionality.

Accessibility

A public digital archive offers accessibility for people all over the world through the Web.

A standard and open technology ensures wide acceptance.

Proprietary technology offers higher flexibility with a risk of nonpersistence and narrow acceptance.

Public Digital Archives

Closed digital archives, such as ones inside museums, have limited outreach. # Public digital archives should be widely accessible from any people, any place using the World Wide Web technology. # Here comes the issue of intellectual property rights, or IPR!!

Copyright

You must check the copyright first!! # Expire after the death of copyright holders such as authors, translators. # The period: Japan 50 years, USA or EU 70 years -- country dependent. # Strong motivation for extending the period (Sonny Bono Copyright Term Extension Act, 1998)

Protect or Share?

Creative work is automatically copyrighted.

Copyright holders cannot give up their copyright, but give permission for other people to use their work.
Media companies try to protect; while people in academia and some artists have motivations to share their work.

Creative Commons ttp://creativecommons.org/

Offering your work under a Creative Commons license does not mean giving up your copyright. It means offering some of your rights to any taker, and only on certain conditions.

Specific conditions for sharing can be explicitly expressed in terms of attribution, derivative work, etc.

2. Input Devices

Digital archive projects start with the digitization of real books.

Compare several types of input devices for the nature of the project.

* The rapid advancement of technologies has made the devices affordable, but digitization still needs some level of expertise.

Popular Input Devices

Scanners

- Flatbed scanners, Film scanners, Drum scanners, Laser scanners, etc.
- # Cameras
 - Analog cameras
 - Digital Cameras
- Robots (Automated digitization)
 High throughput digitization

Scanners

Flatbed scanner

- Commercial products are inexpensive, yet high performance.
- A perfect device for small and unbound documents and books.

#Film scanner

- Useful for digitizing photographs (slides) taken previously.
- Maximum 4000dpi, very high resolution.

Cameras

Digital Camera Technology is improving quite fast because of the high volume of sales. Professional digital cameras are comparable to professional analog ones. # Analog Camera Nowadays only being used when you need both slides and digital images.

Professional Digital Cameras



Kodak DCS Pro



Power Phase FX JASS'04 , July 14th 24th 2004, Tokyo, NII

Single-shot type
4,536x3,024 pixels
36bits color (12bits per color)

Scanner type # 10,500x12,000 pixels # 42bits color (14bits per color)

Robots



... The Future of Publishing

www.4digitalbooks.com # Digitize 250 pages in 10 minutes.

How to Take Pictures?



Image Courtesy Horiuchi Color LTD.

How to Take Pictures?



Image Courtesy Horiuchi Color LTD.

Problems

BIOGRAPHICAL NOTICES.

to him. Of the more comprehensive ecclesiastical biographies containing notices of Odoric, I mention only a few of the chief.

 The FIRST PRINTED EDITION of the Himerary was published by Pontico Virunio, of Belluno, in 4to, at Pesaro in 1513, under the title of Odorichus de Rebus Incognitis. It was printed by Girolamo Soncino, who had presses also at Rimini, Fano, and Orthona ad Mare, from a MS, which Pontico obtained from one Francesco Olivieri of Jesi. It is in the vernacular, in what Zeno calls lingua inculta e rozza, and which Pontico considered to be Odoric's own language (Apostolo Zeno, Dissortazioni Vossiane, Venez, 1751, ii, p. 257). I have not, I regret to say, been able to find a copy of this work. I suppose it is of extreme rarity.

2. In 1529 was printed at Paris for John St. Denys a small folio in black letter with some woodcuts, containing the collection indicated above as French MS. No. 7500 c., under the title of "L'hystoire Merveilleuse Plaisante et Recreative du Grand Empereur de Tartarie, Seigneur des Tartres, nommé le Grand Can," &c. In this, Odorie's work (in French) occupies form f. 53 terd.



Distortion

False color

#Non-uniform Lighting.

Taking Good Pictures

A high-precision book cradle should be used to make the whole page as flat as possible.

Camera position and orientation should be carefully adjusted to get a less distorted image.

Lighting should be carefully designed to have uniform illumination over the whole page.

Flexibility of Digital Images

The advantage of digital images is flexibility; you can apply modification or refinement later when appropriate. # Simplified image acquisition with careful post-processing may reduce the total cost of digitization. # Analog (microfilms) data are still cheaper, but digital data will be cheap.

Large Books and Documents

Image mosaicking is used to make a single picture from a piece of pictures recording a part of target. # Image distortion due to the process of image acquisition requires careful treatment on image mosaicking. # Image mosaicking is a technically established method.

Other Lights

#Laser Scanner

The 3D structure of books can be measured by means of laser scanners. The "contents" of books is more important than their "3D structure." **#X-Ray or Multispectral Sensors** Sensors other than visible wavelength can capture "hidden" information. Sometimes it's worth recording other wavelengths.

Summary

If books are not rare and could be unbound, flatbed scanners are most efficient.

If the scale of digitization project is more than several million pages, the introduction of robots pays.

Otherwise, manual digitization with digital cameras is the best choice.

3. Image Processing and Color Management

Images are the most basic data source in the digital archive of rare books - they have all information. # Images should be processed to extract information from them and use them for other purposes. # Quality control is an important concept throughout the workflow.
Image Formats

#Lossless Coding TIFF (Tag Image File Format): Widely accepted on various platforms. PNG (Portable Network Graphics) : Most Web browsers support this format. #Lossy Coding JPEG (Joint Photographic Experts Group): Good compression rate, and acceptable quality; the most popular.

Lossless Coding

After encoding and decoding, the image can be recovered perfectly. # Compression rate is small or zero. # Suitable for graphic drawings, but not suitable for pictures taken by cameras due to small variations. # The original image * must* be archived with this coding!!

Lossy Coding

Images cannot be recovered perfectly after encoding/decoding. # But the compression rate is high! # Compression rate can be controlled by the quality parameter. # Best for natural images, because small variations can be omitted without visually significant difference.

Problems of JPEG Coding

JPEG PNG

¹ In preparing thi Lee's abridgement, t reading of it. If I ca in Lee which do not plete work. This is # The JPEG image shows artificial noise

around characters, while PNG doesn't.

42,

Color Management

Color is usually represented by three numbers (color space).

This is related to human visual system that has three types of color sensors.

The visual appearance of color is dependent on the input/output device.
Hence "device-independent" color is required.

Device Independent Color



Color Management System

ICC (International Color Consortium) defines the standard for color management.

 # A standard color space is sRGB.
 # For high quality image processing, color management is required all the way from the input to the output.

Image Processing

Control the quality of the image

- Image enhancement
- Noise removal
- #Extract information from the image
 - Image analysis

Character recognition (Addressed later)
 An example of quality control : Digital Bleaching

Digital Bleaching Result

Jaballaha, who lian Christians, be inhabited by

or the bones of Ienezes in 1522, ircumstantially king whom he building of the emoved to Goa. at Thomas was Jaballaha, who lian Christians, e inhabited by

or the bones of fenezes in 1522, ircumstantially king whom he building of the emoved to Goa. at Thomas was an Nach in time to th ir mount to t the famil i ister) wi to have six fi H nost noble ind whose

Decolorization

JASS'04, July 14th 24th 2004, Tokyo, NII

Bleaching

Digital Bleaching

How to improve text readability by means of image processing? #Gray-scale images give an impression of photographs, not books. # Digital bleaching improves text readability by bleaching an image in a sense similar to bleaching wood pulp to make white paper.

Summary

Images are the primary data source in the digital archive of rare books.
Quality control is an important issue throughout the workflow of archiving and preservation.
Image analysis is an advanced issue,

and need to be explored in the domain.

4. Metadata and Annotation

For the evolution of the digital archive of rare books, we need additional information on the data. # Metadata offers description of data for organizing and searching them. # Annotation refers to people's contribution to adding related information about data.

Metadata

Author, the year of publication, description, etc. are the typical metadata of books.

Extracting full text is very useful as additional information of images.

Extracting image features from drawings and pictures inside scanned images maybe useful (not dealt here).

Dublin Core <u>ttp://dublincore.org/</u>

Interoperable online metadata standards, and agreeable baseline metadata for the wide range of materials.

Creator, date, description, identifier, language, publisher, rights, subject, title, etc. are minimum metadata as the information of books.

Lots of Lots of Metadata.

Domain-specific metadata are now actively and competitively developed. Getty Standards : Art and Architecture Thesaurus (AAT) Text Encoding Initiative (TEI) International Standard Archival Description (ISAD) # How can we combine them?

Electronic Texts (etexts)

Manual Production

- More accurate done by human experts.
- Cost may be prohibitive.
- # Automatic Production
 - Optical Character Recognition : OCR

Accuracy is not guaranteed, but is fast and efficient. A good choice for the scale of digitizing a whole book.

Manual Production of Etexts

People type in text.
More reliable than automatic one, but proofreading is still necessary.
Production cost may be prohibitive.
One solution is collaborative work by hundreds of volunteers.

Project Gutenberg <u>ttp://www.gutenberg.org/</u>

- # Michael Hart (U. of IL) started in July 1971.
- Type in books whose copyright had expired.
- As of Oct 2003, more than 10,000 books, 1,000 volunteers, 350 new books per month.
 Plan: a million books by 2015.

Automatic Production of Etexts

Optical character recognition (OCR) has been developed for tens of years.
Accuracy is more than 99% for wellscanned images.

 # Many commercial software packages are available, such as OmniPage and FineReader, for multilingual OCR.
 # Asian languages are still challenging.

The Limitation of OCR

Variation of scanning conditions such as tilt and distortion greatly affects the performance of the result.

Small characters and accents are difficult to recognize.

Dictionary-based correction scheme doesn't work for unknown words, that are often found in rare books.

Search Inside the Book <u>ttp://www.amazon.com/</u>

#120,000 books, 33 million pages of searchable text (Oct 2003). # The company cleared copyright issues by agreements with publishing companies for promotion of sales. # By means of OCR (?), full-text search of recent books is possible, which greatly enhances searchability.

Annotation

Annotation is the process to give additional information.

It includes not only metadata that describe data, but also additional contents that enrich data.

Annotation naturally has a hyperlink structure which is in good accordance with the World Wide Web.

Open Ontology Forge (OOF)

Content annotation and ontology building tool.

Annotate a web page that contains multimedia content (image and text).

Coreference annotation is the unique aspect of this tool.

Freely available under GNU GPL.

Ontology

The explicit and formal specification of conceptualization. Taxonomic hierarchies. Vocabulary, terminology # Knowledge sharing and reuse in the framework of Semantic Web. # Problem: define taxonomic hierarchies for cultural artifacts.

Coreference

Coreference is a relation which refers to the same thing. # Relation between two linguistic expressions (for example, the full name and its abbreviation). # Our extension: relation between a linguistic expression and an image region can be annotated.

Screenshot (1)



Screenshot (2)



Screenshot (3)



Summary

Both metadata and annotation are important for the evolution of the digital archive.

#Richer metadata and annotation will lead to richer knowledge base.

* Many experts' contribution is indispensable for creating the rich collection of metadata and annotaion.

5. Database Systems

Database systems facilitates searching into the digital archive.
On registration, the data should be transformed according to the database schema.

Query language is designed for a user to interact with database systems.

Relational Databases

Author	Title	Year	ID	
Le Coq	Buddhism	1922	LFB-2	
Yule	Cathay	1866	III-2-F- b-2	

Data model is a relation (table).
Most popular database systems now.
Most popular query language is SQL.

SQL Examples

SELECT author FROM author_table WHERE birth_year < 1900; Select authors from "author table" whose year of birth is before 1900. # SELECT * FROM book_table WHERE title LIKE 'A%' ORDER BY sales DESC; Select all columns from book_table whose title begins with 'A' and sort retrieved books by the

decreasing order of their sales.

Full Text Search Engine

One of the most popular interface for searching information (eg. Google). # The result of OCR is used for creating the index (with OCR errors). # Searchability is enhanced significantly with this functionality. #Natural language processing is required for improving performance.

Generic Engine for Transposable Association (GETA)

This engine can be directly used to realize associative searching systems.
Accept a group of texts as queries, and return highly related texts in the relevance order.

Combining with natural language processing, this engine can work as a full text search engine.

Starting a Search

	🚰 Digital Archive of Toyo Bunko Rare Books : Text Search - Microsoft Internet Explorer
	」 ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H) (日)
2	アドレス型 @ http://dsr.nii.ac.jp/toyobunko/search/index.html.en
í.	Digital Archive of Toyo Bunko Rare Books
	Digital Silk Roads Project > Toyo Bunko Archive > Text Search
1	ja en
	:: Text Search ::
	Buddha
	Powered by OFFIN Generic Engine for Transposable Association (GETA) has been developed under the auspices of Dokusouteki Jouhougijyutu Ikusei Jigyo (Innovative Information Technology Incubation Project) of IPA (Information Technology Promotion Agency of Japan).
	Text Search Map Search Project Disclaimer Site Map Contact News
2	Copyright (C) 2003-2004 National Institute of Informatics and The Toyo Bunko. All Rights Reserved.
1	
1	
	الا المراجع الم

Getting a Search Result

	1 675		nttp://dsr.nii.ac	.jp/ogi=bin/toyobunko/geta_search.pi:iang=en&sn=aii&input=budona	<u> 、</u> (* 1>単)
			D	igital Archive of Toyo Bunko Ra	re Books
Di	<u>gital</u>	<u>Silk R</u>	oads Project	t > Toyo Bunko Archive	
	Resi	ults	of Term Se	earch : 1-30 of 186	
				:: Related Documents ::	:: Related Words ::
	1	23	4 <u>56</u> 7		Term Search Reset
		Ass	sociative S	Gearch Reset	
	4			1FB-2 > V-1	1. 🗖 <u>buddha</u>
	1			Die buddhistische Spätantike in Mittelasien Die Plastik / Page 2 (0.000)	
	2			LFB-2 > V-1 Die buddhistische Spätantike in Mittelasien Die Plastik / Page	Term Search Reset
				<u>11</u> (0.000)	
	3			LFB-2 > V-1 Die buddhistische Spätantike in Mittelagien zu Die Plastik (Dage	
			1	<u>12</u> (0.000)	
	4			LFB-2 > V-1 Die buddbistische Spätantike in Mittelagien → Die Plastik (Dage	
			17	<u>16</u> (0.000)	
	5			LFB-2 > V-1	
Continuing an Associative

Search

/ 1.		El http://dsr.nii.ac.	p/cgr-bin/toyobunko/geta_assoc.pl?lang=en&sn=all&dselect=9		∑ ♂移動
		D	igital Archive of Toyo Bunko Ra	are Books	
Digit	tal Silk	Roads Project	> Toyo Bunko Archive		
Re	esults	s of Associat	tive Search : 1-30 of 401		
:: Related Documents ::				:: Related Words ::	
_					
	<mark>1</mark> 23	8456789	2 10 11 12 13 14	Term Search	Reset
	A	ssociative S	earch Reset		
Г				1. 🗖 <u>abgelegene</u>	<u>ם</u>
			Die buddhistische Spätantike in Mittelasien Die Plastik /	2. 🗖 <u>abgeschlag</u> e 3. 🗖 analog	<u>ene</u>
			<u>Page 9</u> (6.532)	4. 🗖 <u>antike</u>	
	2		LFB-2 > V-1	6. 🗖 <u>antiker</u>	
	~		Die buddhistische Spätantike in Mittelasien Die Plastik /	7. 🗖 <u>anziehen</u>	
			Page 18 (0.924)	9. asiens	
	з Г		LFB-2 > <u>V-1</u>	10. ausgegrabe	nen m
		4.73	Die buddhistische Spätantike in Mittelasien Die Plastik / Page 12 (0.902)	12. ausgewählt	<u>er</u>
		Distant and		13. 🗖 <u>ausmaß</u> 14. 🗖 ausmaßen	
	4		LFB-2 > V-1 Die buddhistische Cententike in Mittelegien – Die Dischlief	15. 🗖 <u>ausspreche</u>	
			Die buddhistische Spatantike in Mittelasien Die Plastik / Page 11 (0.894)	16. Daustein 17. Doi	
				18. 🗖 <u>bediente</u>	
	5 🗖		LFB-2 > V-2	19. <u>bildhauerei</u>	

Natural Language Processing

Part-of-speech tagging # Stemming # Stop word elimination # Frequency calculation # Indexing structure # Search by keywords # Search by contents

Asian Language Processing

Relatively well developed for Japanese, Korean, and Chinese. # Most of the languages are harder to analyze than European languages. #Natural language processing requires a large-scale software development, hence requires a long time devotion to research activities.

Summary

The above database engines mainly concerns the retrieval of data, but the database has more possibilities, such as mining information. # The design of new database models and new query languages is still a hot topic of research. What is the best model for the digital archive?

6. World Wide Web and Case Studies

- # As a public digital archive, having a web site is a must.
- * The information design of the web site is briefly introduced.
- # Several considerations characterize the direction of development.
- # The digital archive developed by the author will be used as a case study.

Open or Proprietary Technology?

Open Technology Specification is open to the public. Widely accessible from the public. Technology tends to survive. # Proprietary Technology Special functionality not available on open technology. If the company goes bankrupt....

Information Architecture

How users interact with the digital archive?

A typical interface is a keyword search. Suppose you see a blank box upon an access. How do you feel?
The system should provide enough navigational methods that reflect user's possible browsing paths.

Simple Navigation Model



A typical interface on the digital archive web site. # Only effective when you know what you want, and the answer is either correct or wrong.

Actual Navigation Model



Content Management System

Content Management System (CMS) is required to maintain an operational digital archive web site.

 CMS supports building and publishing the web site, but existing systems generally do not directly support creating a digital archive site.
 Case study: built from scratch.

Digital Archive of Toyo Bunko Rare Books

The digital archiving project of rare books started in 2002, as a part of Digital Silk Roads Project. # Collaborator: The Toyo Bunko. # As of Jul. 2004, 27 titles with 5623 pages, taken by digital cameras. # The system was built on Linux with numerous open source software.

Content Management



#XML (eXtensible Markup Language) is intensively used for the management of data in the web site. **#XSLT(XML Stylesheet** Transform) is used for transforming XML to HTML.

Appearance

Digital Arc Digital Silk Roads Project > Toyo Bunko Archivi - Bibliography Page Index High Res Page J Color Mid. Res. Image Gray High Ro Previous Die buddh	chive of Toyo Bunko Rare Books e > LFB-2 > V-1 > Page View Index Facing Pages Index Thumbnail Index Caption Index es. Image ja	en •xt
Digital Silk Roads Project > Toyo Bunko Archiv - Bibliography Page Index High Res Page I Color Mid. Res. Image Gray High R Previous Dia buddh	e > <u>LFB-2</u> > <u>V-1</u> > <u>Page View</u> Index Facing Pages Index Thumbnail Index Caption Index - es. Image ja	en ext
Color Mid. Res. Image Gray High R Previous Die buddb	es. Image ja	en ext
Previous Die buddh	N	<u>ext</u>
Die buddh		
0093	istische Spätantike in Mittelasien Die Plastik Postancient Buddhist Culture in Central Asia Sculpt	ures

14 C (1 15 15	14 . 5 *	1
Na	VIOO	TIO	1

Digital Silk Roads Project > Toyo Bunko Archive > LFB-2 > V-1 > Page View

- Bibliography -- Page Index -- High Res Page Index -- Facing Pages Index -- Thumbnail Index -- Caption Index -

Color Mid. Res. Image

Gray High Res. Image

Previous

Hierarchical structure.
Sibling information.
Alternative representations (resolution, language, etc.)
Sequential access (prev, next page).

JASS'04, July 14th 24th 2004, Tokyo, NII

<u>1a</u>

Next

en

Other Navigation Paths

🔽 🔗移動 | 📆

ja en

Digital Archive of Toyo Bunko Rare Books : Map Search - Microsoft Internet Explore ファイル() 編集() 表示() お気に入り(A) ツール() ヘルブ(H) アドレス() () http://dsr.nii.ac.jp/toyobunko/map/index.htmlen

Digital Archive of Toyo Bunko Rare Books

Digital Silk Roads Project > <u>Toyo Bunko Archive</u> > <u>Map Search</u>



Browsing through geographical location. # Browsing through temporal information. # Other browsing modes

Having Good Relationship with Popular Search Engines

#Let search engine robots come in to the digital archive and index the contents to make them searchable. # The database should be organized to be visible from crawling agents, by generating static pages, for example. # People can then use Google, etc. to directly find pages of interest.

Summary

The design of a digital archive web site requires several design decisions in the beginning.

System designers should consider information architecture from the viewpoint of users.

Efficient content management system can reduce the maintenance overhead.

7. Future Direction

Strong trend toward the large-scale digital archiving of valuable materials. # The technological advancement of input devices, storage devices and others support the construction of large-scale digital archives. # Collaboration among motivated people will become more and more important.

Future Direction

Copyright (intellectual property rights) is getting more strictly protected, which may adversely affect open digital archives. # Commercial and non-commercial digital archives should take appropriate roles to deliver cultural heritage to people in need.

Resources

- # Digital Silk Roads Project
 - http://dsr.nii.ac.jp/
- # Digital Archive of Toyo Bunko Rare Books
 - http://dsr.nii.ac.jp/toyobunko/
- # Open Ontology Forge (OOF)
 - <u>http://research.nii.ac.jp/~collier/resources/O</u> <u>OF/index.htm</u>
- # GETA (Generic Engine for Transposable Association)
 - <u>http://geta.ex.nii.ac.jp/</u>

KITAMOTO Asanobu

- # Associate Professor, National Institute of Informatics
 - More information at <u>http://www.kitamoto.org/</u>
- KITAMOTO Asanobu graduated from University of Tokyo in 1992, and earned Ph.D. from University of Tokyo in 1997. Since then, he works for National Institute of Informatics (and its predecessor, National Center for Science Information Systems) in the areas of image processing, pattern recognition, image database systems, data mining, and remote sensing. In 2001, he joined the digital archive project "Digital Silk Roads Project" and has been working for the digitization of rare books. He and other members now run a public Web site "Digital Archive of Toyo Bunko Rare Books" (<u>http://dsr.nii.ac.jp/toyobunko/</u>), on which 27 rare books can be viewed from the cover to the bottom with full-text search and other navigational assistance.

