Context Recombination for Digital Cultural Archives

Asanobu KITAMOTO, Sonoko SATO, Takeo YAMAMOTO, and Kinji ONO

National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{kitamoto, sonoko, ty, ono}@nii.ac.jp
http://dsr.nii.ac.jp/

Abstract. We first introduce our digital archive called ``Toyo Bunko Digital Archive." The target of this digital archive is Silk Roads-related rare books written in European languages, and now the archive has more than 5600 images from 27 rare books. The digitization of those books is performed using professional digital cameras, following the automatic processes to obtain electronic text using optical character recognition (OCR) and machine translation to improve accessibility to the contents of book pages. Then we propose the concept of "context" as the key strength of the digital cultural archive, and illustrate some examples on context recombination, which is the main topic of this paper. Then we break down this concept into technical challenges, and it leads to the development of context-based database systems and the design of a new query language, which will be used for our digital cultural archive in the next generation.

1. Introduction

Our research group is involved in the digital archive project called ``Digital Silk Roads (DSR)," which is a joint work with UNESCO under Digital Silk Roads Initiative Framework (DSRIF). The purpose of the DSR project is to propose a new approach for the preservation and utilization of the large amount of Silk Roads-related cultural heritage based on collaborative work between information technology and cultural studies.

This paper begins with the introduction of a DSR project called ``Toyo Bunko Digital Archive." The target of this digital archive is rare books related to Silk Roads. In our sense, the digital archive means more than just the simple scanning of book pages. It moreover aims at building a virtual place in which interaction and communication over the digital archive leads to the enhancement of our knowledge through the discovery of new interpretations and understandings. The central concept for designing such a digital archive is what we call "context."

Then the last half of this paper deals with discussion on what we mean by context. Context is, roughly speaking, any kind of relationships among a group of data. Then we propose a simple model that links the context with the context; that is, we represent the content as the combination of two elements -- context-sensitive content and context-free content. A few examples then illustrate examples of how we can recombine various contexts, or context recombination, for exploring the space of information to discover new interpretations and understandings. Then we discuss the technical challenge of context recombination, namely the context-based database system and the design of its query language, which leads to the redesign of traditional database systems.

2. Case Study: Toyo Bunko Digital Archive

2.1 Background

National Institute of Informatics (NII) made an agreement with UNESCO on the Digital Silk Roads Initiative Framework (DSRIF) in 2001. 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 large amount of cultural heritage based on collaborative work between informatics and cultural studies. The role of informatics communities is to provide information infrastructure to renovate cultural studies on Silk Roads. The DSR project members consist of researchers from more than ten countries including many central Asian countries, and several research programs have already started among members. More detailed information is available on the portal site of the DSR project[1].

Among several research programs, our project, Toyo Bunko Digital Archive, focuses on the digital archive of rare books related to Silk Roads. This 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, many of the books are “invisible” from the general public because of limited access to rare books mainly for preservation and safety purposes. Therefore, the improvement of accessibility to those rare books is one of the goal of this project.

2.2 Goals

Our goal is not limited to the building of an “old-fashioned” digital archive that simply provides still images of rare books. From the perspective informatics, we propose that the digital cultural archive is not only a repository place for storing and maintaining digital cultural resources but also a gathering place for stimulating interaction and communication between motivated people over digital cultural resources for scholarly research and education activities.

The former is the traditional notion of the digital archive, which is simply the replacement of real storage rooms with the digital space combining both digitization techniques of real cultural artifacts and simulation techniques of the real world. We, however, mainly focus on the latter, namely sharing interest and enjoying with digital

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

cultural artifacts. Technical challenges toward this ambitious goal include a systematic support for enhancing the amount of knowledge, and the redesign of traditional database systems from the viewpoint of context recombination, which will be introduced later.

We also have more practical goals to improve accessibility to rare books with reference to technological innovations in related areas of books and Web as follows.

1. The full text search service of published books provided at Amazon.com[2]. They use optical character recognition (OCR) to extract textual information from published books. With the full text indexing, all pages in books are searchable by keywords. This service has dramatically improved accessibility to the contents of books compared to the traditional style of searching books using simple metadata like the title and the author.
2. The machine translation service of the Web documents provided by SYSTRAN[3] and others. The accuracy of machine translation is still not satisfactory, but it is better than nothing. Now many Web documents are written in other languages than English, and, with caution, even translated text with many errors serves to be the useful source of information.

These two services describe some of the desirable features of the digital archive. We suggest that the full text search and machine translation will be the integral part of the digital archive in the near future.

To match those services, we applied optical character recognition (OCR) to extract textual information from digitized books, and then machine translation and image processing for the automatic analysis of digitized documents.

2.3 Selection

Among Toyo Bunko 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 books in Morrison Library, and initiated the digitization of rare books in FY 2002. In two years we digitized 5,623 pages from 27 rare books[4], ranging from the report of academic explorations to more personal travelogues.

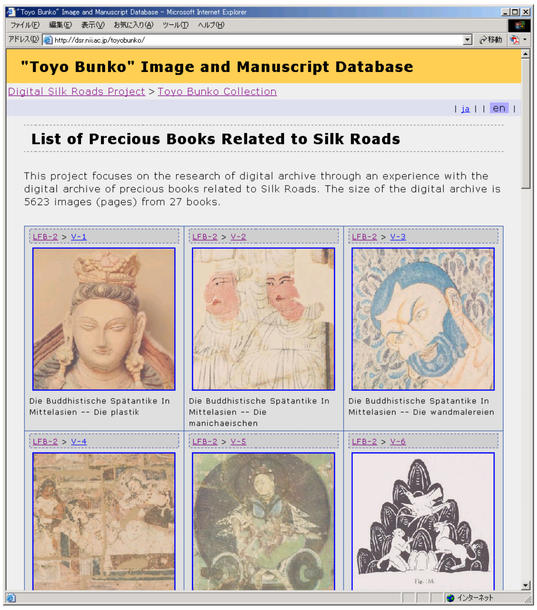
The special value of these books originates in the lively description of people, customs, cultural artifacts, geological features, historical events, spoken or written languages and transcribed texts which has been changed, destroyed or lost since 19th and early 20th century. In terms of languages, those books are written in several European languages (French, Italian, German, English and Russian).

[2] http://www.amazon.com/

[3] http://www.systransoft.com/

[4] All the books are free from copyright restrictions.

Another unique aspect of those books is that they contain many illustrations and photographs. These illustrative books have richer contents compared to text-only books in terms of information that is hard to describe by words. Nevertheless, the importance of images, photographs and illustrations has been overlooked in the research communities due to the historical importance of text resources. We therefore plan to build a digital archive with visual information.


"Toyo Bunko" Image and Manuscript Database - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://dsr.nii.ac.jp/toyobunko/
移動
インターネット

"Toyo Bunko" Image and Manuscript Database

Digital Silk Roads Project > Toyo Bunko Collection

| ja | | en |

List of Precious Books Related to Silk Roads

This project focuses on the research of digital archive through an experience with the digital archive of precious books related to Silk Roads. The size of the digital archive is 5623 images (pages) from 27 books.

LFB-2 > V-1
Die Buddhistische Spätantike In Mittelasien -- Die plastik

LFB-2 > V-2
Die Buddhistische Spätantike In Mittelasien -- Die manichaeischen

LFB-2 > V-3
Die Buddhistische Spätantike In Mittelasien -- Die wandmalereien

LFB-2 > V-4

LFB-2 > V-5

LFB-2 > V-6

Fig. 1 Toyo Bunko Image and Manuscript Database.

2.4 Digitization

Cover-to-cover digitization is performed using high-quality professional digital cameras with the resolution of either 10,500 x 12,600 pixels or 4,072 x 4,072 pixels. Books were placed on a book cradle that supports the material without applying severe stress to the binding.

Then we applied commercial OCR software packages to obtain electronic text. The motivation behind the use of automatic processes like OCR is our need to manage the large number of books. The automatic approach is possible partly because our selection of books is limited to those written in major European languages, neither in Asian languages with complex character sets nor in lost languages with hand-written or cursive scripts. Among several software packages tested, we found that the performance is almost comparable with the error rate of less than 1 percent with a small variation dependent on scanning conditions and written languages. [1]

The performance of OCR is satisfactory for documents with a good digitization condition and a low layout complexity. However, the performance is degraded significantly due to bursty recognition errors for documents with the presence of tilt, footnotes, figures and captions. One source of error comes from the complex layout of old books because of their manual designs by typesetting experts. Even a single page may contain multiple fonts with different sizes and typefaces, especially in captions and footnotes. Another source of errors comes from the presence of accent signs in some European languages. We also noticed that the evolution of languages and the frequent occurrence of geographical names with notational variations may be another reason of degraded performance in spite of the usage of dictionary-based correction methods.

2.5 Accessibility

The portal site of Toyo Bunko Digital Archive is shown in Fig. 1. The web site is almost completed but will be open after non-technical issues are solved. The main content of the portal site is Toyo Bunko Image and Manuscript Database[5]. This database provides the digitized images of rare books in the maximum resolution of 900 x 900 pixels. However, we are planning to increase the resolution so that even small characters on the page are readable on the screen. On the portal site, navigational links for book browsing and a full text search engine is provided so that a user can find a page with the keyword and directly jump to the page with a digitized image and an OCR text.

3. Content and Context

3.1 What is the real strength of the digital archive?

What is the real strength of the digital archive? A typical answer to this question is the preservation of real cultural resources. It is true that the digital form of cultural resources is relatively safe from degradation and destruction, but at the same time, it also has the longevity problem of format and media compatibility. Another typical answer is the improvement of accessibility by means of searching and networking technology. It is true that digital cultural resources are easy to find with keywords and easy to access with the Internet, but at the same time, it is the improvement of efficiency but not qualitatively innovative. Hence we argue that the real strength of digital cultural archives is beyond preservation and accessibility. The real strength is interaction and stimulation.

Interaction refers to the communication and collaboration over the digital archive among the communities of people. The idea of collaboration between motivated people has attracted a lot of attention after the success of open source movement, and some people actually started applying this principle into cultural domains. The idea itself is attractive, but in practical applications, it requires breakthrough in the management of information quality, which we will not discuss in this paper.

Stimulation refers to the discovery of new understandings and interpretations of cultural resources. Our solution is to offer multiple viewpoints systematically based on the advanced usage of the databases, or data mining. But this process is fundamentally an interactive process between humans and machines to find a good alternative to satisfy users' information needs against the inherent ambiguity of cultural artifacts.

Here comes the word "context." We insist that the recombination of context can be used as a unifying principle to formalize the process over digital cultural archives that may lead to the discovery of new interpretations and understandings of cultural

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

artifacts. For the better definition of the concept we begin with discussion on the comparison of context in the real space and the digital space.

3.2 Exhibition and Context

The presentation of cultural artifacts requires the organization of a collection in a systematic manner. An exhibition is an example of the organization of a collection by curators so that the organization visualizes their ideas and concepts in the form of the intended arrangement of cultural artifacts in a real space under various constraints such as space, location, environment and collection. We regard an exhibition as a kind activity toward the creation of a context in the sense that a visitor's mind is inspired and stimulated by the curator's intended context, under which the visitor subsequently wants to interpret and understand the collection.

This example demonstrates the interesting aspect of the collection of cultural artifacts. It suggests that even a collection of limited size may inspire people with new interpretations and understandings when different arrangements are carefully tried out. We hence focus on this fundamental nature of an exhibition, namely an activity to create a meaningful context from the collection, or an activity to recombine the context, to take advantage of the inherent ambiguity of cultural artifacts.

Similar arguments also apply to libraries, where the arrangement of books reflects a perspective on the organizations of books in librarians' mind. We can say that the arrangement (or exhibition) of books is an activity to create a meaningful context in which books are searched and browsed by visitors. Since the discovery of new books is often made by coincidence among books placed nearby on the same shelf, for example, librarian's activity to create context plays an important role in the discovery of knowledge from a book collection. A book collection is usually organized by topic, but obviously this is not the only context that a librarian can create.

The activity of exhibition can gain more freedom in the digital space. The arrangement of cultural artifacts is now free from the constraint of real space of museums and libraries. Now we can set up the arrangement of cultural artifacts in multiple ways to offer visitors richer contexts, or even we can rearrange the collection instantly on visitors' request. In short, the digital archive allows us the liquid design of organizing the collection. The collection itself is now free from the collection of a single museum and library and can be extended to the inter-museum or inter-library sharing of digital cultural resources with a standard protocol and agreement. Finally, the exhibition is free from spatial and temporal constraints and any person in the world can visit the exhibition through the Internet.

These examples indicate that traditionally context is created in the form of exhibitions, but in the digital space, we can exploit the variety of possible approaches toward the creation of context. We however need a systematic support for experimenting different recombination of context, and technically speaking, this amounts to the design of a database system, or more specifically, the design of a query language that supports the recombination of contexts at user's information need. We will discuss this technical challenge in Section 4.

3.3 The Model of Content

Context can be defined in relation to content. Here content represents information contained in a resource. Roughly speaking, context can be defined as any kind of relationships among the contents of a group of resources. The context so defined is regarded as the content of the group. More formally, we assume that the content of a resource can be represented by the combination of two basic elements, namely context-free content and context-sensitive content as follows.

Content = **Context-free Content** + **Context-sensitive Content** (1)
Context-sensitive Content = **Relationship** + **Arrangement** (2)

The content defined as Equation (1) can be recursively defined as the context-free content of a group of resources, and we can assume higher level context-sensitive content among a group of the groups. Equation (2) represents that context-sensitive content is further decomposed into two basic elements, namely relationship and arrangement. We then discuss the role and the purpose of those basic elements.

3.4 Context-free Content

Context-free content is based on the hypothesis that the content of a resource can be analyzed independent of other resources. The contents of a resource are extracted automatically or semi-automatically as features of the resource using various image and text processing algorithms, and especially in terms of visual features, many researchers tried to extract contents that cannot be expressed by simple keywords, such as color, layout, and impression. Only this aspect of the content has been considered in most of the content-based multimedia database systems [2,3].

The success of content-based multimedia database systems, however, is still remained primitive to this day in spite of enormous effort in this area. It is argued that the reason is the shallow level of analysis for extracting semantic information from multimedia signals. Nevertheless, little is known about the feasible direction of research toward deep semantic analysis, and it seems that a breakthrough, or the change of thinking, is necessary in this area.

We argue that the omission of context is another reason of limited effectiveness. It is now clear from previous arguments that the semantic interpretation of a resource is fundamentally dependent in which context the resource is interpreted and understood. Suppose we want to represent the semantic information of a painting. The painting can be interpreted in the context of the life of the painter, in the context of the group of the painter, in the context of a method of painting, and in the context of the historical evolution of painting. Hence the semantic interpretation of a painting is meaningless without the specification of context you want to interpret and understand the particular painting.

Hence the contents of a resource cannot be determined with their own right, but depend on the context in which a resource is related to other resources in terms of similarity, dissimilarity, commonality, semantic relationships, and others. This part of content is what we call the context-sensitive content.

3.5 Context-sensitive Content

Context-sensitive content is based on the hypothesis that the semantic interpretation of a resource requires the specification of a context. This hypothesis may be supported by the well-known theory of linguistics stating that the meaning of a concept cannot be defined in isolation, but can be defined in relation to other concepts. Following the same argument, we can say that the content of a resource is defined in relation to other resources.

We also assume that we can further decompose the context-sensitive content into two basic elements, namely relationship and arrangement. Figuratively speaking, relationship takes part of librarian's work for the organization of information, while arrangement takes part of curator's work for the presentation of information. Naturally those two roles overlap each other, but there still remains some difference in the fundamental roles of their work.

3.5.1 Relationship

Relationship takes the various form of connection among the collection of resources. An example of relationship is similarity, where the degree of relationship is computed according to a specified distance measure. If the distance measure has controllable parameters, the change of those parameters may lead to a new set of relationships between resources. Another type of relationship is based on the metadata of resources. For example, we can define commonality relationships between resources that share the same keyword, or the concept that the keyword implies. More complex type of relationships is based on hypertext and hypermedia, where the network of annotations on resources or other annotations represent the rich network of information. For the logical design of relationship we can borrow well-known models from the field of information design, such as hierarchy, table, and categorization. These logical structures with the organized representation of semantic relationship can be understood easily by people in general and hence are effective as the model of relationship.

3.5.2 Arrangement

Arrangement is required both in the visualization of relationship and the presentation of context. The basic operation of arrangement is the ordering of resources based on the value of the content or the metadata. The ordering gives contextual information in terms of the position of the value in the domain or the cumulative probability distribution. That is one of contextual information in the sense that we can obtain information about whether the value is an extreme value or a normal value. The ordering combined with winnowing (collecting) is also effective

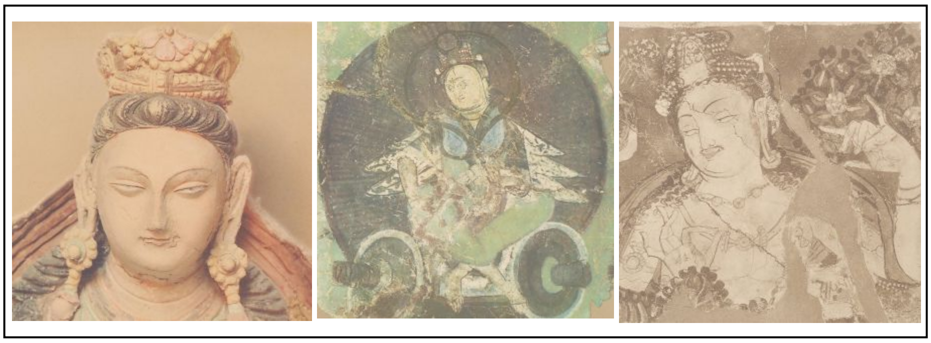


Fig. 2 Context Recombination by Serial Arrangement.

for creating the top-N group of resources. If necessary, ordering can be performed in a higher dimensional space to visualize relationships in a higher dimensional space. When the relationship is represented by hypertext, arrangement involves the extraction of a subgraph, which is then serialized to form the group of resources with a possible subsequent ordering.

4. Context Recombination

4.1 Basic Ideas

In Section 2, we discussed the model of content as the combination of context-free content and context-sensitive content. The model of context-free content has been studied extensively in the field of content-based multimedia database systems, but context-sensitive content have been less studied, and we still do not have prescriptions as to what kind of context is useful for the specific information need of users. Hence the system should support the flexible recombination of context with the explicit specification of how to recombine the context. The repeated process of context recombination may lead to the creation of meaningful contexts within which new interpretations and understandings are discovered. In the following, we present a few specific examples about the recombination of contexts.

4.2 Context Recombination by Keywords

Keywords are given to each cultural resource as a part of metadata of the resource. It may be a controlled keyword or a word in a free-text description. If you ask a database engine to retrieve resources that contain some keywords, you will obtain a group of resources that share the same keywords. This process can be described, from another viewpoint, as obtaining a set of resources that share the same features or the same concepts that the keywords represent. Hence this group can be regarded as a meaningful context, because what is not shared among members suggests semantic

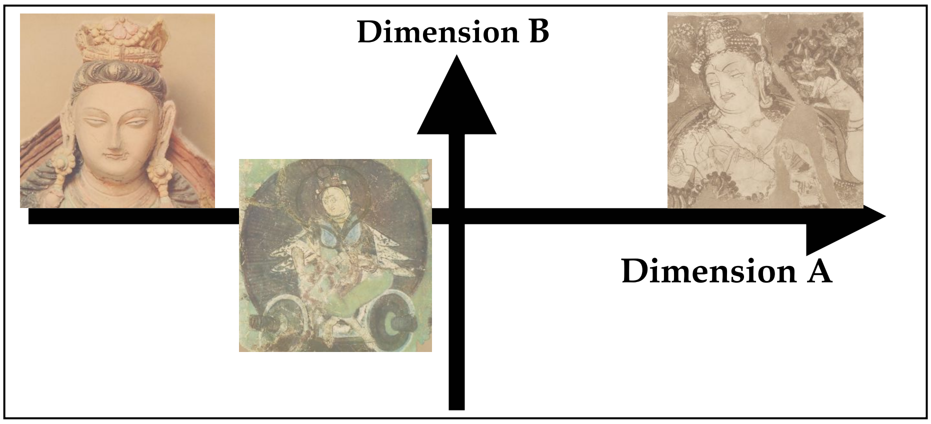


Fig. 3 Context Recombination by Spatial Arrangement.

information about each resource. The subsequent ordering of cultural resources based on the time of creation, for example, leads to another meaningful context in which the evolution of style can be discussed. The grouping of resources based on artists and schools also creates a context, as is often done with real exhibitions.

4.3 Context Recombination by Serial Arrangement

A simple serial arrangement of Buddha images as in Fig. 2 creates a meaningful context in the same sense as an exhibition does in a real space. This type of context is closely related to the human cognitive process. It has a tendency toward recognizing stronger relationship between spatially closer objects than objects that are far away. It is related to the fact that humans try to read meaningful information even from noisy signals with an expectation that something could be found in the signal. Finally it is also related to the nature of human memory that, when a serial arrangement is viewed sequentially, information seen in the near past has more influence than that seen in the far past. These tendencies of the human cognitive process should be exploited to stimulate the creativity of humans toward innovative interpretations. Thus the arrangement of resources is directly linked to the recombination of contexts, and the rearrangement of the same collection of resources may lead to other contexts with appropriate proximate and ordering relationships.

4.4 Context Recombination by Spatial Arrangement

The arrangement of resources may be performed in a higher dimensional space to take advantage of the degree of freedom in a higher dimensional space. For example, we can place resources on a two-dimensional space as shown in Fig. 3 so that the distance between resources give information about the degree of relationship between resources. The degree of relationship is usually measured as the distance between resources in terms of the similarity of image features or semantic contents. The

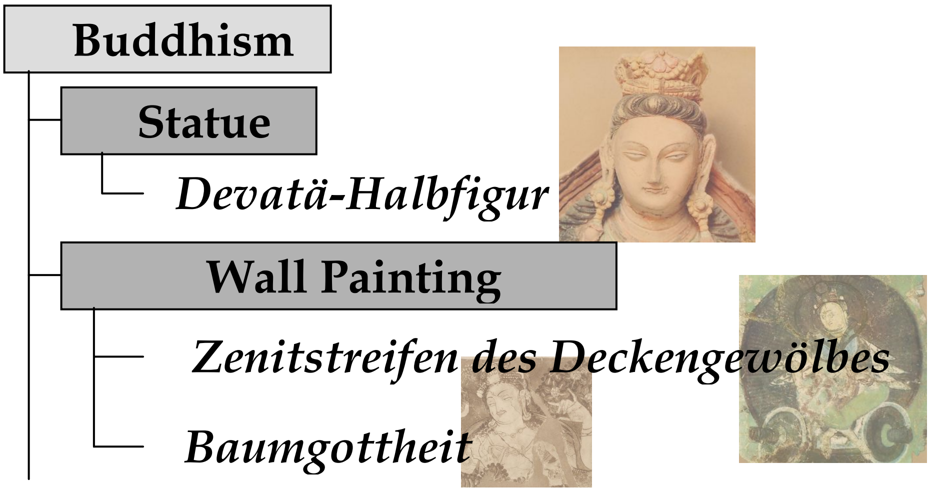


Fig. 4 Context Recombination by Taxonomy (Ontology) Tree.

presentation of relationships in a higher dimensional space gives intuitive ideas about relationships based on multiple viewpoints.

4.5 Context Recombination by Taxonomy (Ontology) Tree

A cultural resource is many-sided and multilateral because of its inherent ambiguity. A cultural resource can be associated with a taxonomy tree in terms of the place, age, creator, method and so on. A subtree in the taxonomy tree as shown in Fig. 4 is the basis of context, since the entries under the parent entry share some features or concepts that are represented in the parent entry. Serializing the subtree into a group can create a meaningful context, in which ordering or arranging can be applied subsequently. Hence taxonomy is a useful device for creating contexts, and the appropriate choice of entries from a taxonomy tree results in multiple contexts for a single resource.

4.6 Context Recombination by Annotation Graph

The final example of the recombination of context is based on the graph of annotations. If we allow an annotation on a resource, an annotation on an annotation, or an annotation on a context, then we have the graph of annotations that link related resources, annotations, and contexts. These links entail some semantic information, so a traversal on the graph of links along paths can create a meaningful context that has recursive relationships. An example shown in Fig. 5 illustrates the case of free-text annotations. An annotator focuses on the shape of the hand of Buddha, and place an annotation like “The shape of the hand is interesting.” If another person finds this annotation, it may trigger the interest of another person into the shape of the hand and

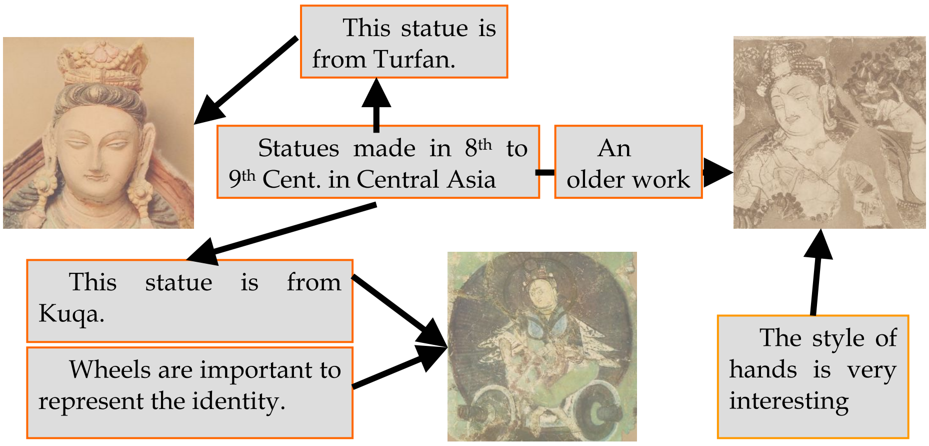

This statue is from Turfan.
Statues made in 8th to 9th Cent. in Central Asia
An older work
This statue is from Kuqa.
Wheels are important to represent the identity.
The style of hands is very interesting

Fig. 5 Context Recombination by Annotation Graph.

follows links to and from the resource about interesting hands. If we have other paintings of Buddha that share the same annotation, then we can create a context about Buddha's hand in which the detail of the difference can be argued and clarified. Hence an annotation can create a context about semantic information and relationships. A similar hypertext structure has been used for a long time, but we use this hypertext structure as a device to create a context, as is the same with a taxonomy tree.

5. Context-based Database Systems

5.1 Framework

Each example in Section 4 seems to be a matter of course. Nevertheless we want to emphasize that our purpose of illustrating these examples is to demonstrate that the recombination of contexts can rule over all examples as a unifying principle. In this framework, technical issues amount to the design of a database system and a query language that supports context recombination. This support is out of the functionality of relational database systems, which are basically context-free content-based database systems. We require the database system to be designed as context-sensitive content-based, and this requires the complete rethinking of the model and the design of database engines. We therefore give another name to this idea, namely the context-based database system.

The purpose of the context-based database system is to provide a context in response to a user's query. This action itself is similar to traditional database systems to the point that the response is a set of resources. Nevertheless, the response from the context-based database system is always an ordered (or more generally arranged) set in comparison to an unordered set from traditional database systems. We even see an ordering or an arrangement in a randomly shuffled set of resources because of the presence of relationships such as proximity. Similar behaviors can be achieved in traditional database systems using operators like "order-by" but those operations are considered as ad-hoc supplements for post-processing the output, so they are not powerful enough for our purpose. More flexible database systems like object-oriented databases could implement our framework, but the flexibility leads to the lack of a principled design like relational database systems, and hence not advantageous.

Among various issues for this modification, the most technically challenging and fruitful area of research is the design of a query language that supports the idea of the context-based database system. The design of the query language is directly linked to the flexibility and effectiveness of the database systems, and is the foundation of deeper technical challenges such as query formulation, query optimization, and indexing schemes.

5.2. Query Language

A query language is a tool to specify the information need of a user and tell it to the database engine in a form that the database engine can parse and process easily. Information need of a user is assumed to follow, in this paper, the model of content proposed in Section 3.3, and the goal of the query language is to support at least basic operations for the recombination of context as summarized in Section 4.

The design of a query language is required to be efficient. The first desirable property is the closure property. In this case, we force that an operation takes a group of resource as the input, and output a group of resources, which can be used as another input to subsequent operations. This closure property is the key to the flexible combination of operations, in particular recursive combinations. Another desirable property is the orthogonality of basic operations. We should have a set of basic operations in which any operations cannot be represented by the combination of other operations. We propose here that the following operators are orthogonal, minimum set of operators to realize the recombination of contexts as introduced. [5]

1. Grouping
   This is the abstraction of grouping, stratification, clustering, and edge traversing operations.
2. Ordering
   This is the abstraction of sorting, rearranging, projecting, and random shuffling.
3. Attribute Expansion
   The list of attribute is expanded to allow volatile attributes.
4. Collecting
   This is the abstraction of winnowing, sampling, and summarizing operations.

We plan to implement those operations in the database engine and helps users to create contexts on the fly. With efficient database operations we aim at proposing an inspiring tool for people to compare digital cultural resources and discover new interpretations and understandings. The preliminary version of the database engine with the partial support of these operators is already working on the Web site "Digital

Typhoon"[6]. The next version of the database engine that fully supports the proposed query language is still under development, but it will be used for our digital archive, namely Toyo Bunko Digital Archive as introduced in Section 2.

5.3. Related Topics

Other topics related to the design of the database system are briefly described here because of the limited space of this paper.

The first topic is information visualization. The database system takes part in the logical aspect of information, while information visualization takes the visual part of information, which is indispensable for clarifying and enhancing the given context with graphical representations. We plan to implement information visualization modules on top of the database engine as a presentation layer in the system hierarchy.

The second topic is metadata. Although we briefly address the usage of metadata, we did not discuss which metadata we will use for our project. We are planning to use widely accepted metadata such as Dublin Core [4]. The specific choice of metadata formats, however, is not a central topic in this paper. Especially in cultural domains we already have many sets of metadata defined by domain experts, some of which are usable for our purpose.

The third topic is scalability. This issue is not well studied in this paper because of our focus on the design of a new database system rather than speeding up an existing database system. Standard indexing schemes such as tree structures for strings or high dimensional vectors can be applied to our framework, but at this moment we simply rely on linear scanning. This is partly because in high dimensional spaces, it is well known that the indexing structure is less effective than that in lower dimensional spaces, and hence the linear scanning is a good alternative. Hence in the future we combine simple indexing structures and linear scanning with pruning.

The fourth topic is performance evaluation. Our system is practically efficient because all the index data can be fit into main memory. Suppose we have 1 million entries and each entry takes 1 Kbytes as the index of the entry. This is a fairly large digital archive at this moment, but it still fits into the main memory of 1 Gbytes which is affordable on today's personal workstations. Hence except for extremely large digital archive projects exceeding 1 million entries, we can assume that the index data can be fit into the main memory, and the performance is acceptable without sophisticated I/O management.

7. Conclusion

The digital archive is not only the repository of digital resources but also the place for the communication and interaction over digital resources. People in the world can visit the Web site through the Internet, leaving some comments or discussing with other people. In this sense, we admit that we focus more on the accessibility aspect

[6] http://www.digital-typhoon.org/

than preservation aspect. It is sad to see that many precious books are left untouched in the dark repository room of the library. The information technology can improve the accessibility to those hidden invisible books for everyone in the world. To realize rich interaction with the databases, especially in the domain of cultural resources, we claim that the notion of context is an important concept. The importance of context has been addressed by many authors, especially the critics of art such as Walter Benjamin. Our argument about context shares some perspective with those previous arguments, but the major difference is a system point of view for the interactive design of context. This viewpoint is especially important in the digital archive, since it enhances our experience with the digital archive, and it enables us to explore the space of digital cultural resources even we are in the house or any place in the world. We illustrated some examples and design principles based on an ongoing work. Context-based database systems require the fundamental redesign of the database systems, so the research on the design of such databases is itself an interesting topic of research.

References

[1] S. Sato, A. Kitamoto, Y. Shiba, I. Tanaka, K. Ono, and T. Yamamoto. Multilingual DSR Document Archive: Digitization, Automatic Multilingual Indexing and Collaborative Thesaurus Construction, In *Proc. Nara Symposium for Digital Silk Roads*, (in press), 2003.

[2]M. Flickner, et.al. Query by Image and Video Content: The QBIC System, *IEEE Computer,* Vol. 28, No. 9, pp. 23-32, 1995.

[3]A. W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain. Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence* , Vol.22, No. 12, pp. 1349-1380, 2000.

[4]A. Kitamoto. Context Recombination Engine (CORE): A Design Concept for Digital Archives. In *Workshop on Digital Libraries for Cultural Preservation.* (in press), 2003.

[5] Dublin Core, Dublin Core Metadata Initiative, http://www.dublincore.org/