Reformulation of Contexts:
A Design Concept for the Database of DSR Archive

Asanobu Kitamoto, Takeo Yamamoto, Sonoko Sato, and Kinji Ono

National Institute of Informatics

kitamoto@nii.ac.jp, ty@nii.ac.jp, sonoko@nii.ac.jp, ono@nii.ac.jp

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

Abstract

We claim in this paper that "context" is a key design concept for stimulating people's mind to discover new interpretations and understandings in the digital archive. Our approach to the notion of context is unique in terms of the proposal of the model of context from the viewpoint of database systems and informatics. We assume that content is actually comprised by two basic elements, namely context-free content and context-sensitive content, of which the latter is usually neglected in traditional database systems. We further propose that the context-sensitive part can be further decomposed into two elements, namely relationship and arrangement. Then we discuss some examples about how we can reformulate the context using various operations, and this arguments leads to the new line of database design, namely context-based database systems with the new design of a query language. Finally we introduce our ongoing digital archiving project called Toyo Bunko Image and Manuscript Database, and introduce some technical challenges in this digital archive with future plans.

Keywords

Cultural Digital Archive, Reformulation of Context, Context-based Database Systems, Query Language, Content-based Multimedia Retrieval, Optical Character Recognition

1. Introduction

The masterpieces of paintings, sculptures and other cultural artifacts always stimulate our mind with rich source of information, emotion and inspiration. Whenever we see them, we may find new interpretations and understandings. After learning art, history and related studies, we can derive even more insightful interpretations and understandings because of the background knowledge we acquire from learning. This example indicates that the value of cultural artifacts, or cultural resources in general, is in its inherent ambiguity and complexity. In the digital archive of cultural artifacts, we therefore should provide a systematic mechanism to take advantage of the ambiguity of cultural artifacts and support the discovery of multiple interpretations and understandings. We propose in this paper that this mechanism can be realized through a systematic support for the flexible reformulation of viewpoints, or contexts, in which cultural artifacts are to be interpreted and understood.

Here context is the keyword of this paper. In our view, context is something that suggests any relationship among resources. The notion of context is similar to the notion of viewpoint, but it not only represents a viewpoint but also the design of information. This paper claims that the context is a key design concept both in the design and implementation of the digital archive, especially in the cultural domain.

The context itself has been a focus of attention in many publications and exhibitions in cultural studies, but little has been proposed on the systematic design of context from a viewpoint of database systems and information infrastructure, which play far more important roles in the digital archive. The contribution of this paper is therefore in the redesign of contexts as the unifying concept of the digital archive, and in the formalization of the context from the viewpoint of informatics as in Section 2 with examples in Section 3.

Then we briefly introduce the new line of database systems, namely the context-based database system. The goal of this database system is to support the reformulation of contexts with a technically sound basis. Technical challenges not only include the design of database systems, but also the design of a query language. Comparison of our ideas with the content-based database systems clarifies the purpose of the context-based database systems in Section 4.

In addition to conceptual work above, this paper also introduces the digital archive of rare books called Toyo Bunko Image and Manuscript Database, which is one of the digital archives under the Digital Silk Roads (DSR) project. The

purpose of this archive is the digital archiving of rare books related to Silk Roads beyond the simple digitization of book pages. Section 5 introduces the detail of this project and future plans. Finally Section 6 concludes the paper.

2. Context and Content

2.1 Exhibition and Context

The presentation of cultural artifacts requires the organization of the collection in a systematic manner. What is called an exhibition is an activity for curators to visualize 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 the same collection may inspire people with different interpretations and understandings from different arrangements, or contexts. In other words, even the limited size of the collection is a source of fruitful ideas through the rearrangement of the collection with an unknown viewpoint. We hence focus on this fundamental nature of an exhibition, namely an activity to create a meaningful context from the collection, which is based on the inherent ambiguity of cultural artifacts.

Similar arguments also apply to libraries, where the arrangement of books reflects a viewpoint for 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. Since the discovery of new books is often made by coincidence by the proximity of books, for example, the creation of context by a librarian plays an important role in the discovery and enhancement of knowledge from a book collection.

The activity of exhibition, however, could gain more freedom in the digital space. This is because the arrangement of cultural artifacts is now free from the constraint of real space of museums and libraries. The arrangement of the collection can be set up in multiple ways to provide visitors with richer contexts, or even the collection can be rearranged instantly on visitors' request. The collection itself is free from the collection of a single museum and library but can be extended over multiple museums and libraries with a standard protocol and agreements over inter-organization sharing of digital archives. 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 the context is traditionally created in the form of exhibitions, but in the digital archive, we should take advantage of the freedom of the digital space, and exploit the variety of possible approaches toward the creation of context in the digital space. A systematic support for the creation of context is indispensable in the digital space, and technically speaking, this amounts to the design of the database system that supports the reformulation of contexts, or more specifically, the design of the query language for the specification of the information need of users about the reformulation of contexts. We will discuss this technical challenge in Section 4.

2.2 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 can also be used as the content of a group of resources. More formally, we assume that content of a resource can be represented by the combination of two basic elements, namely context-free content and context-sensitive content as follows.

$$\underline{\text{Content}} = \underline{\text{Context-free Content}} + \underline{\text{Context-sensitive Content}} \quad (1)$$

$$\underline{\text{Context-sensitive Content}} = \underline{\text{Relationship}} + \underline{\text{Arrangement}} \quad (2)$$

The content obtained in the 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 a group of resources. Above equations also represent that context-sensitive content is further decomposed into two basic elements, namely relationship and arrangement. In the following, we will discuss the role and the purpose of those basic elements.

2.3. Context-free Content

Context-free content is based on the hypothesis that the content of a resource can be analyzed independent of other resources. Only this aspect of the content is usually considered in content-based multimedia database systems [2], in which the

contents of a resource are extracted automatically or semi-automatically as features by means of various image and text processing algorithms. Using those methods, many researchers have been trying to extract visual contents that cannot be expressed by simple keywords, such as color, layout, and impression.

In spite of enormous effort in this area, however, the success of content-based multimedia database systems is still remained primitive to this day. It is argued that the reason of this failure is the shallow level of analysis methods for representing semantic information contained in a multimedia signal. Nevertheless, little is known about the feasible direction of research toward deep semantic analysis, and it seems we are required to find a breakthrough in this area.

We argue that the omission of context is another source of inherent limitation of the success. It is clear from arguments above 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.

It is now clear that 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.

2.4. 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 in which it is argued 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 largely overlap each other, but there still remains some difference in the roles of their work.

2.5. 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 [3], where the network of annotations on resources or other annotations represent the rich network of information. For the 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 easily understood by people and hence are effective as the model of relationship.

2.6. Arrangement

Arrangement is required both for 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 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.

3. Reformulation of Contexts

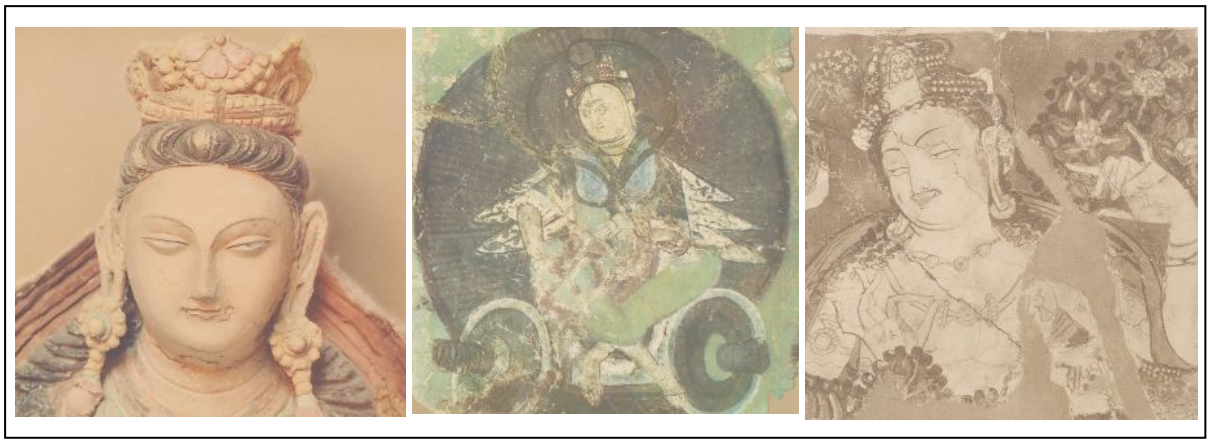

Fig. 1 Reformulation of Contexts by Serial Arrangement.

3.1 Basic Idea

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 focused, hence 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 flexibly support the various form of information need on how to reformulate the context with the explicit specification of how to combine basic operations. The specification is described by the query language, and the reformulation of contexts is performed in the database engine on request. The repeated process of the reformulation of contexts then leads to the discovery of meaningful contexts within which new interpretations and understandings are discovered.

In the following, we present a few specific examples about the reformulation of contexts. We plan to implement the following operations in the database engine and the query language. With this framework we aim at providing an inspiring tool for people to compare and to discover the relationships among resources with the help of contexts created on the fly.

3.2 Reformulation of Contexts 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 the 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 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.

3.3 Reformulation of Contexts by Serial Arrangement

A simple serial arrangement of Buddha images as in Fig. 1 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 reformulation of contexts, and the rearrangement of the same collection of resources may lead to other contexts with appropriate proximate and ordering relationships.

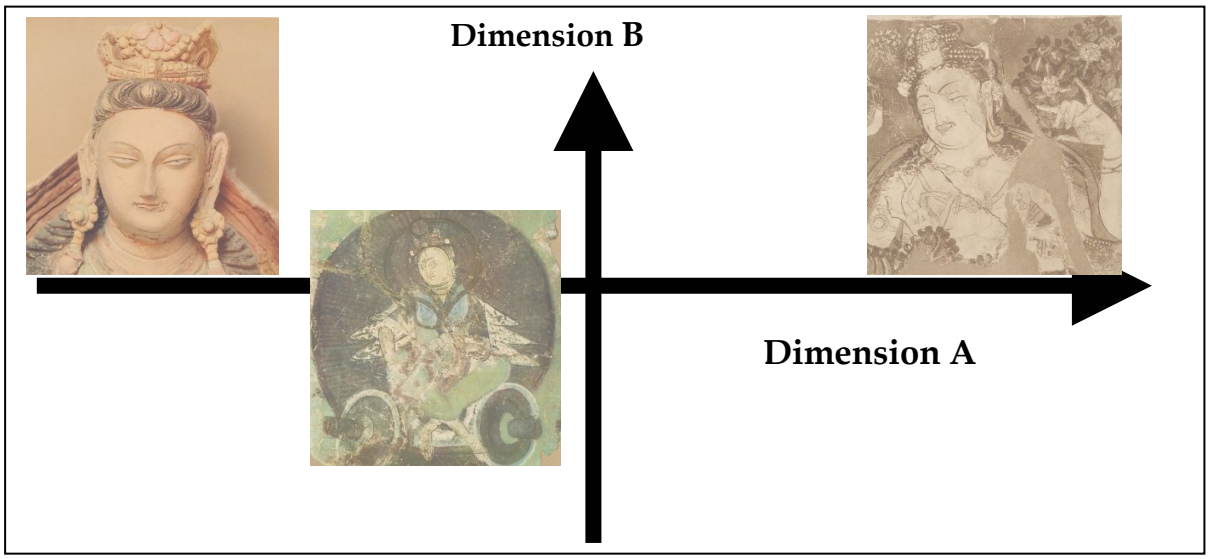

Fig. 2 Reformulation of Contexts by Spatial Arrangement.

3.4 Reformulation of Contexts 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. 2 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 presentation of relationships in a higher dimensional space gives intuitive ideas about relationships based on multiple viewpoints.

3.5 Reformulation of Contexts 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. 3 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

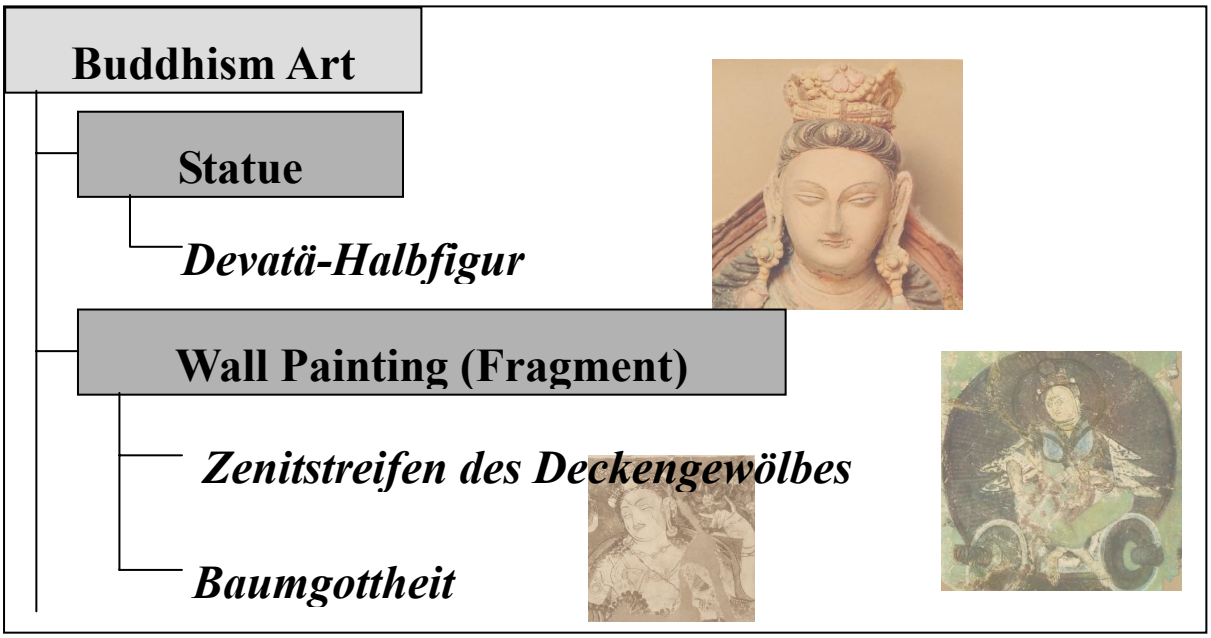

Fig. 3 Reformulation of Contexts by Taxonomy (Ontology) Tree.

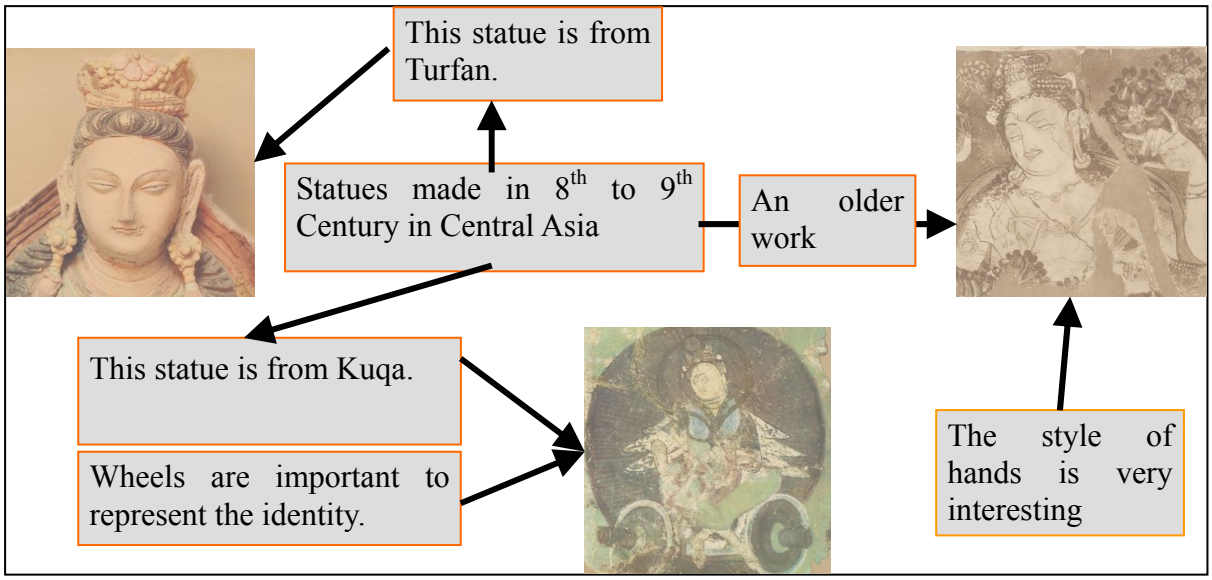

Fig. 4 Reformulation of Contexts by Annotation Graph.

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.

3.6 Reformulation of Contexts by Annotation Graph

The final example of the reformulation 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 traversing on the graph of links along paths can create a meaningful context that has recursive relationships. An example shown in Fig. 4 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 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.

4. Context-based Database Systems

4.1 Basic Idea

Each example in Section 3 seems to be a matter of course. Nevertheless we want to emphasize that our purpose of illustrating these examples is to demonstrate that the reformulation of contexts can rule over all examples as a unifying concept. Based on this observation, we now turn to the technical design of the context, namely the design of database systems and query language based on the reformulation of contexts. From a database point of view, this idea requires the way of change in thinking, because the new role of the database is to find a context, which is different from finding a resource in the traditional database design.

We therefore give another name to this database system, namely the context-based database system, which is named after the content-based database system. As discussed before, the purpose of this database system is to provide a context in response to a query. This action 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. This kind of new formulation requires considerable change of thinking on the design of database systems.

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 with 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.

4.2. 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 2.2, and the goal of the query language is to support at least basic operations for the reformulation of context as summarized in Section 3.

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 reformulation of contexts as introduced. [4]

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.

The preliminary version of the database engine that supports a part of the proposed query language is already working on the Web site "Digital Typhoon"[1]. The next version of the database engine that fully supports the proposed query language is still under development, but on the completion. it will be used for our digital archive, namely Toyo Bunko Portal Site, as will be introduced in Section 5.

4.3. Related Topics

Other topics related to the design of the database are briefly described here. Although these topics are necessary in the whole design of the digital archive, they are not integrated into the database system itself.

The first topic is information visualization. The database system takes the logical part of information, while information visualization takes the visual part of information, which is indispensable for enhancing the effect of context. Information visualization receives logical relationship between resources from database systems and outputs the rendering of relationships in a graphical manner. We plan to implement this module 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 in fact planning to use widely accepted metadata such as Dublin Core [5]. The specific choice of metadata formats, however, is not an important argument in this paper. Especially in cultural applications, metadata formats are defined by many domain experts, some of which are usable for our purpose.

5. Toyo Bunko Image and Manuscript Database

5.1 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 huge amount of cultural heritage by means of cooperative work between informatics communities and cultural

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

studies communities. The role of informatics communities is to provide efficient 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 detailed information about these projects is available on the portal site of the DSR project[2].

Among several research programs under the DSR project, our project, Toyo Bunko Project, 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, 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.

5.2 Goals

Our goal in this digital archive is not limited to the building of an "old-fashioned" digital archive that simply provides still images of rare books, but is extended to the demonstration of novel approaches for the design of the digital archive from the viewpoint of informatics. We also propose that the digital archive is not only a place for storing digital resources but also a place for stimulating interaction and communication between digital resources and the community of interested people.

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 by means of virtual reality, for example. We, however, mainly focus on the latter, namely the model of the digital archive in logical aspects. We aim to design the digital archive so that it can support not only the maintenance of digital cultural resources but also the stimulation of cultural research and education activities systematically for enhancing the amount of knowledge in the digital archive. The challenges in this direction are how we can provide a place for people who want to study, communicate, and collaborate over the digital archive.

Aside from these ambitious goals, we also have more practical goals to improve accessibility to rare books. Toward this goal, we can refer 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[3]. 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. The improvement of accessibility from this service is just enormous compared to traditional style of searching books by a simple metadata.
2. The machine translation service of the Web documents provided by SYSTRAN[4] 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.

5.3 Book Selection and Digitization

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 5,623 pages from 27 rare books, 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,

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

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

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

those books are written in several European languages (French, Italian, German, English and Russian). Another unique aspect of those books is that they contain many illustrations and photographs. Compared to text-only books, these illustrative books have richer contents, but historically the importance of images, photos and illustrations has been overlooked in the research communities. We therefore plan to build a multimedia digital archive with the help of visual information processing techniques. Finally all those books are free from copyright restrictions.

The 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.

5.4. Text Processing

For text processing we applied commercial OCR software packages. The motivation behind the employment of automatic processes like OCR is the need to manage the large number of books. This 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 (1) the performance is almost comparable with the error rate of less than 1 percent, (2) OmniPage Pro 12 Office (ScanSoft) seemed to be relatively stable against the variation of scanning conditions and written languages. [1]

The performance 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 rare books. They have more complex layout than today's books due to manual designs by typesetting experts. Even a single page may contain multiple fonts with different sizes and typefaces, especially at captions and footnotes. Another source of errors comes from the presence of accent signs in some European languages. We also point out that the evolution of language 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.

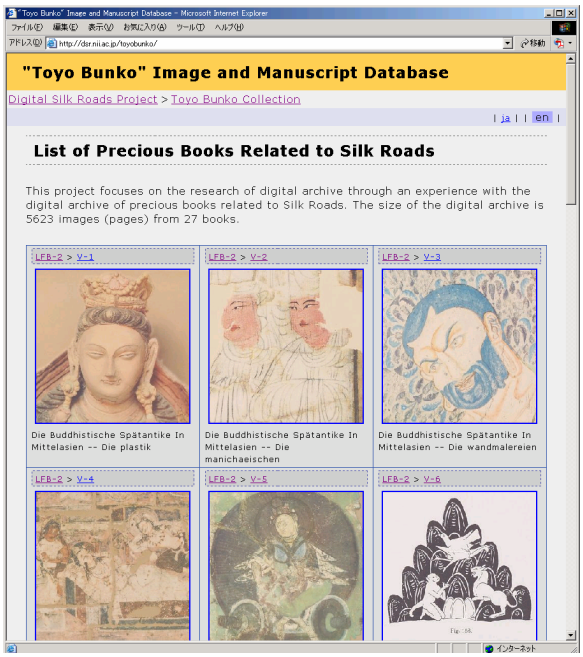

Fig. 6 Toyo Bunko Image and Manuscript Database.

5.5. Portal Site

The portal site of Toyo Bunko Project is now ready for opening as shown in Fig. 6. The main content of the portal site is Toyo Bunko Image and Manuscript Database[5]. This database serves the digitized images of rare books with the maximum resolution of 900 x 900 pixels. We, however, are planning to increase the resolution for display so that even small characters on the page are readable on the screen. In addition to navigational links for book browsing, we also provide the full text search of books so that a user can find a page with the keyword and directly jump to the page to study the digitized image and the OCR text in the same Web document. The detail can be found elsewhere. [6]

5.6. Future Plans

The system addressed above is only the first version in our roadmap. We are planning to implement the following systems in the next step to enhance the functionality of the digital archive.

1. Multimedia Data Processing System
   This system is used for extracting meaningful information from multimedia elements such as illustrations, photographs and maps. This is necessary for searching pages based on the content of multimedia data, such as the color

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

distribution of a photograph in a page.

2. Annotation Management System
   In our ambitious goal, the collective annotations by the community of users are the integral part of the digital archive. The annotation may include the combination of a free text comment and a formal metadata, and linked with other resources or annotations. For this purpose, we need a formal management system for the hypertext annotations.

3. Context-based Database System
   We already introduced the concept of the context-based database system, but it is yet to be fully implemented, tested and used. Hence the implementation should be listed in the top priority, and should be tested on the digital archive, such as Toyo Bunko Image and Manuscript Database.

6 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. Information technology can support those goals, and improve the accessibility to rare but invisible books for everyone in the world, ultimately leading to the enhancement of our knowledge.

As a design concept for such information infrastructure, we claim that the notion of context plays a fundamental role for taking advantage of the inherent ambiguity of cultural artifacts. We illustrated some examples of context reformulation, and proposed the context-based database system and its query language to support the reformulation of contexts. This database system is based on a different concept from traditional database systems, so the design of such databases is itself an interesting topic of research.

Toyo Bunko Image and Manuscript Database is our first effort on the digital archive. Although it is just a first step, this is a relevant step toward the sharing of cultural information and the discovery of knowledge contained in precious cultural artifacts, which are waiting for our access in the dark rooms of libraries and museums.

Acknowledgment

The authors thank Dr. Yoshinobu Shiba and Dr. Issei Tanaka in the Toyo Bunko for their helpful support to the digitization of rare books in the Toyo Bunko.

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, this volume, 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] J. Conklin. Hypertext: An Introduction and Survey. IEEE Computer, 1987.

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

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

[6] A. Kitamoto, E. Platon, F. Andres, and T. Yamamoto. NII Portal Sites for the Digital Silk Roads Project, In Proc. Nara Symposium for Digital Silk Roads, this volume, 2003.