US – Korea Joint Workshop on Digital Libraries
August 10-11, 2000
San Diego Supercomputer Center
San Diego, California

1. Executive Summary

There are many barriers to the worldwide development of digital libraries, and the involvement of the US in such activities. These are of particular concern in the context of digital library support of collaboration on research and education between pairs of nations with very different languages and cultures, such as the US and Korea. We report on recommendations to remove such barriers that were developed through a workshop involving digital library researchers from the US and Korea who met August 10-11, 2000 at the San Diego Supercomputer Center.

Attendees contributed position statements prior to the workshop.

2. Introduction

2.1. Background
2.2. Motivation
2.2.1. Mutual Rewards from US / Korea Collaboration
2.2.1.1. Opportunity for Improving Cross-Cultural Understanding
2.2.2. Strong Strategic Ties between US and Korea
2.2.3. Strong Academic Relationships
2.2.3.1. *Koreans with US academic affiliations (numbers?)*
2.2.4. Strong Economic Relationships
2.3. Regional Focus / Global Implications
2.4. Emerging Collaborative Activities
2.4.1. The Pacific Rim Digital Library Alliance

Digital libraries provide a mechanism to assemble collections that span multiple countries, enabling access to material that would otherwise be inaccessible. An example of such a collaboration is the Pacific Rim Digital Library Alliance (PRDLA). The members of the alliance constitute institutions from the nations that surround the Pacific Rim, including a Korean university library. Each institution contributes either collections, software infrastructure or hardware support systems to build a distributed digital library. One of the goals is to assemble the largest collection of Chinese text in the world by integrating access to multiple existing collections. The following collections under PRDLA are operational. Note that intellectual property rights restrictions limit access to some collections.

- Bibliography of East Asian Studies provides an index to Western language journal articles on East Asian studies. The index development is led by Stanford University. To access this database, point your browser to prl.sdsc.edu and follow the “Bibliography of East Asian Studies” link.

- Beijing SuperStar Digital Library provides scanned images of some 80,000 titles of Chinese digital books on many subjects from China. Some of the books are mirrored in a repository at the San Diego Supercomputer Center, with the entire collection located in China. The collaboration includes Beijing SuperStar - a major vendor of

Chinese digital books in China. To access this database, point your browser to prl.sdsc.edu and follow the “Beijing SuperStar Digital Library” link.

- The electronic version of Wenyuange Sikuquanshu (The Wenyuan Library collection from the Qing Palace) contains full-text and scanned images of some 3,700 titles. The collaboration includes a vendor of digital books in HongKong – Digital Heritage. Access to this database requires vendor supplied client software that is licensed to each university

The US-Korean joint workshop on digital libraries pointed to the need to extend these collaborations to include cultural information from Korea. The PRDLA shows that the policy and cooperative agreements needed to build a multi-national library can be implemented to further communication between nations.

3. Technical Directions and Challenges for Global Digital Libraries
    3.1. Scalable Content and Collections
        3.1.1. Lessons from Applications in the Humanities
        3.1.2. Globalizing Digital Libraries and Museums
        3.1.3. Globalization of Scientific Research
        3.1.4. Transforming Education
    3.2. Architectural Considerations

The development of a global digital library has unique requirements not present in current distributed digital library projects. While the architecture is substantially similar, the need to support multi-lingual interfaces and multi-lingual knowledge bases poses significant computer science research issues.

The simplest characterization of a digital library environment differentiates between the management environments required for digital objects (data), attributes about the digital objects (information), and relationships between the attributes (knowledge). Each environment requires explicit infrastructure support. Fortunately, storage of data has been facilitated by the development of data handling systems (SDSC Storage Resource Broker), and information management has been enabled by the acceptance of a standard information markup language, XML. Management of knowledge has been a major research area for decades, and the recent integration of XML markup languages with Topic Maps for describing relationships promises to provide a suitable infrastructure. Ontologies can now be expressed as XML tagged relationships that associate topics. It is possible to create representations of ontologies that can be implemented on a variety of logic systems.

The challenge in managing knowledge is compounded by multi-lingual access. The concepts used in one language may not map directly into concepts used in another language. This means even though the same domain knowledge is being described related to the same collection, a multi-lingual digital library will have to maintain multiple ontologies to describe the relationships used to organize the collection. Relevant research questions include whether it is possible to build a joint ontology, whether the mapping between the ontologies can be automated, and whether cross language information retrieval can facilitate the ontology mapping.

The architecture for a global digital library is simplified by considering the following infrastructure levels:

- **Application/Configuration layer**
  This level provides tools to help select digital objects for assembling a new digital library, as well as tools for defining context as concept spaces. Concept spaces are needed for both defining the level of expertise of the user, as well as the level of expertise of the collection.
- **Multi-lingual Digital Library Services**
  Multiple services are provided by a global digital library, including presentation, translation, knowledge mining, information mining and query support. Presentation services include tools for dynamically defining the user context and modifying the presentation context to be compatible. Views are needed to control user specified presentation versus collection specified presentation such as choice of character set. Translation services provide an opportunity to develop correspondences between concept spaces that function at the level of a thesaurus. Knowledge mining provides tools for the identification of relationships. Information mining provides tools to seek existence of explicit relationships, such as content based information retrieval.
- **Domain Knowledge Management**
  Ontologies manage concepts for the collection, and require explicit tools to facilitate their maintenance. Tools are needed to extend an ontology to include new concepts that describe common sets of information about the collection. Tools are also needed to support the inverse task, creation of metadata for concepts that are contained within the ontology, but not expressed within the collection metadata. Tools are also needed to integrate existing ontology representations.
- **Collection Management**
  The ability to build an infrastructure independent description of a metadata catalog is needed to support migration of collections between different types of database systems. Effectively, this consists of tools to support creation of schema, modification of schema, and publishing of schema for use by other applications. The collection can then be distributed across multiple information repositories, with local control over the digital objects, and access control lists used to protect intellectual property. A distributed collection can be implement across multiple administration domains, including internationally distributed sites.
- **Data Handling System**
  When digital objects are distributed across multiple storage systems, data handling environments are needed to manage access. This is essential for integrating collections that are distributed across file systems, archives, and databases provided by multiple vendors. Distributed digital object management provides persistent identifiers, replicas of data sets, containers for aggregating digital objects before storage in archives, and archival storage interfaces.
- **Storage systems**
  The fundamental level of a global digital archive is the set of storage systems used to hold the data objects. The storage systems will be distributed internationally, with each site maintaining local control. A global digital library only becomes possible when the storage systems are integrated by a data handling system.

        3.2.1. Distributed Repositories
            3.2.1.1. Open Archives
            3.2.1.2. Virtual Documents
        3.2.2. Integration and Interoperability
            3.2.2.1. Information Integration Models and Techniques
    3.3. Bridging the Language Barrier
        3.3.1. Machine Translation and Cross-Lingual Information Retrieval
        3.3.2. Multi-lingual User Interfaces
    3.4. User Access
        3.4.1. Content-based Multimedia Access
        3.4.2. User Interface Design
        3.4.3. Mobile Access
4. High Priority Technical Challenges
    4.1. Ubiquitous Accessibility
    4.2. Interoperability

The development of a global digital library can be thought of as the creation of interoperability mechanisms for access to data, information, and knowledge. The competing approach is to rely upon standards for storage, schema definition, and ontology definition. The standards-based approach has been tried for large multi-national data collections, but requires a tremendous effort to reach consensus. The resulting system usually requires use of a single storage system (such as object-oriented databases), a common schema (with pre-specified lists of allowable attributes), and a common ontology (with explicitly defined relationships between the domain concepts and the sets of attributes used in the collections). The challenge is that all aspects of the digital library will evolve in time, with new concepts added to represent new findings, new attributes used to describe new data sets, and new types of storage systems created that are larger and faster. If a standards-based digital library is to persist, all of the components will have to upgraded simultaneously. This will not be feasible in most global digital library efforts.

By building an infrastructure that focuses on interoperability, new interfaces are built whenever technology evolves, and the system can continue to operate with both the original and evolved components. Standards are still important, but at the information markup language and knowledge representation level. The storage systems used, the contents of each collection within the digital library, the schema used to organize the collection, and the ontology used to represent relationships about the collection can be different. The interoperability mechanisms provide the ability to map between the different instantiations.

This approach is also needed to ensure persistence of the global digital library. The mechanisms used to support interoperability in space between heterogeneous systems are the same as those needed to support migration of collections in time onto new technology. The challenge that global digital libraries present is the need to agree on standards for information languages and knowledge representations. But this challenge is precisely what must be faced for information to become available and accessible globally.

4.2.1. Necessary but not sufficient conditions for 6

Possible Extension to Section 4.2 - RBA
Compared to text, much less is known about multimedia indexing and access. MPEG7, the emerging multimedia description standard, should be the foundation for much of this effort. This program should provide opportunities to integrate MPEG7 descriptions into the digital library services and end-user applications. For instance, a multimedia collection server might use MPEG7 for indexing its collection and groups of these multimedia servers could be organized as multimedia open archives.

Advanced services will require developing new multimedia content and actively re-purposing existing content. The current MPEG7 proposals need to be greatly expanded to allow specification of multimedia services such as descriptions of problem sets, classroom lesson plans, and interactive hypermedia presentations of lectures.

.

4.3. Cross-Language Issues
4.3.1. MT
4.3.2. CLIR
4.3.3. Extraction & Summarization
4.3.4. Adaptive Interpretation (tailored presentation to user)

5. High Priority Application Areas
5.1. Leveraging Korean Investment in Cultural Heritage
5.1.1. Working Toward a Vision

Korean cultural heritage provides a rich opportunity for digital library collaboration between the US and Korea. Korea has large collections of historical resources that are only available *in Korea*, and, hence, not available to international researchers without those researchers traveling to Korea. Currently there are efforts within Korea to scan ancient documents for preservation purposes. Precious ancient documents and artifacts are at risk and must be preserved before they deteriorate beyond usability, are lost or destroyed. These efforts are capturing static images, rather than generating truly digital documents. Even those materials that are available digitally rarely conform to international standards that would make them usable by the international community, and researchers that are successful at finding digitized resources still find the language barrier a serious impediment to usage.

Cultural exchange between the US and Korea is out of balance. Koreans tend to be substantially more familiar with US culture and language than US citizens are about Korea, and many more Koreans come to the US for education and employment than US citizens go to Korea. Given the strength of the economic and diplomatic relationships between our two countries, substantial benefit can accrue to both countries through a more balanced exchange. This is clearly inhibited by a range of technical issues, such as standards conformance metadata support, and character and document encoding, and further by a significant (and lop-sided) language barrier. Seemingly straightforward and simple solutions such as the use of Optical Character Recognition (OCR) to convert page images into digital documents have not matured sufficiently to offer viable solutions for Korean materials.

Collaboration between the US and Korea on preserving and providing access to cultural resources offers a mutually rewarding opportunity for advancing the technologies underlying global digital libraries. Such a collaborative endeavor can be composed of three phases: proof of concept, focused development and marketing, and production operations.

It's new & novel!

The first phase (proof of concept) captures the interest of the technology community and the stewards of cultural resources, and strives toward concrete, well-defined objectives to set the stage for longer term expansion. This prototyping period requires the construction of a balanced core corpus of materials in which the breadth and depth of resources is represented, including content such as text, images (of art objects, people, places, rare books, …), maps, dictionaries, and semantic aids such as ontologies. This phase also presents the opportunity to construct 3-D representations of interesting objects and spaces, and to build bilingual resources such as lexicons and parallel corpora.

Phase 1 yields a "best of class" prototype based on the current state of the art, to prepare the way for long-term development and confront fundamental questions in areas such as preservation and access. Its operational goal is to demonstrate feasibility, functionality, and utility. Its pragmatic goal is to establish a critical mass of people, infrastructure, and information resources. Its technical goal is to evaluate the sufficiency of current approaches, techniques, and standards.

It's useful & desirable!

The second phase of development moves from proof of concept to deployment and expansion, bridging multiple disciplines and domains. An explosion of applications, users, and uses is accompanied by a growing recognition of the need for interoperability, metadata standards, ontologies, machine translation tools, and scholarly resources such as commentaries and handcrafted translations. This phase confronts issues of scalability and sustainability directly, validates the architecture and approach, and accommodates new advances from research and development. This is also the phase where user evaluation studies are most relevant and rewarding, as a statistically significant set of resources, applications, and users become available within a coherent domain for experimentation and study.

It's assumed & unnoticed!

Phase 3 may be the invisible phase, characterized by Digital Library resources and facilities being so common and fundamental to educational and scholarly pursuits that few would even consider working without them. This sets the stage for transformation of education and research. Phase 3 requires cross-lingual transparency, or the ability to operate in one's preferred language on a set of materials in other languages with proficiency. Cross-lingual transparency requires effective cross-lingual information retrieval (CLIR), content-based multimedia support, and information extraction and summarization tools. It enables seamless multilingual cross-disciplinary research and analysis, and cross-cultural learning and collaboration, leading to a fundamental transformation of manual scholarly practice.

Phase 3 also leads to documents that are *designed* for a digital library, which is to say that they would not be feasible without the infrastructure of a DL in place. One could easily imagine routinely time-tagging and geo-referencing new digital objects in the collection. Images, for example, would always have a geographic reference and time associated with them, enabling easy correlation of images across time and space. One could also attend to a vast array of details that confound scholarly use of information, such as disambiguating proper names, resolving co-references within and among documents, and establishing authority control over information resources.

Policy

Experience in other digital library projects (e.g., Perseus) has established the need to clarify intellectual property arrangements early, and to avoid compromises and assumptions. Particularly in international projects, where intellectual property laws and mores may differ among the participants, it is essential to reach a full and complete understanding early, to make the obvious explicit, and to establish the principal of monotonic progress (i.e., once something is in the collection, it is never out of the collection).

Capturing the Moment

The time is right for a US / Korean collaboration on a Cultural Heritage Digital Library. While digital library research and development will continue for a long time, the decade or so of development to date has yielded substantial returns, resulting in a mature set of basic digital library technologies. Coupled with the dramatic growth in networking worldwide, and particularly the development of broadband capacity in the US and the extensive wired and wireless infrastructure of Korea, the basic technological infrastructure is in place.

Korea is committed to a massive digitization effort that will yield a wealth of Korean cultural heritage information resources, available for global access and utilization. This effort recognizes the need for conformance with global information standards and best practices, so is sure to provide an invaluable source of materials.

The US, through the National Science Foundation (NSF) is encouraging and facilitating the development of international digital library projects. Cooperative efforts have been established with several European countries, but to date no such relationship has been established with an Asian country. Korea provides a timely and propitious opportunity, with benefits accruing to the global community in the large, but with a special opportunity for the US to gain access to high quality information on Korean culture that is essentially unavailable anywhere in the US today. A joint US / Korean cultural heritage digital library project could also expand the accessibility of American cultural resources to Korea, and enable greatly enhanced multicultural education and collaboration opportunities.

A joint initiative between the US and Korea would necessarily engage researchers in industry and universities in both countries and their government sponsors. Its reach would extend beyond the traditional computer science researchers engaged in digital library work to include cognitive scientists, librarians, and other disciplines.

A successful cultural heritage initiative will require massive sets of digitized information resources produced through government resources (e.g., MIC in Korea), coupled with comparable corpora from existing DL's in the West. The initiative will als require an array of language technologies, both existing and under development. But, perhaps most important, the initiative will need a strong and coherent voice of leadership.

It will need to consider a broad range of performance measures and criteria by which success can be evaluated, including both quantitative and qualitative attributes. Quantitative measures must clearly include size and usage statistics, including the numbers of users, the number of digital objects, the relative proportion of US vs. Korean users, etc. Qualitative measures must also be developed to assess international usability, value, and return on investment.

5.2. Science, Mathematics, Engineering, and Technology (SMET)

5.2.1. Working Toward a Vision

Multiple international science-based projects are being driven by the need to globally generate, access, and analyze data. The projects span the "hard" sciences, including

- High-energy physics – the CERN large hadron collider will generate petabytes of data per year starting in 2005, which will be analyzed by researchers in Europe and the US.
- Astronomy – a Virtual Observatory is being proposed that integrates digital image data from all-sky surveys generated from telescopes in both hemispheres.
- Biology – the Protein Data Bank distributes protein structures through international mirror sites
- Earth systems science – remote satellite sensor data is generated by multiple nations, and integrated through data assimilation to provide both weather forecasts and serve as primary data for investigating global warming.
- Neuroscience – the Human Brain Project builds upon research at multiple institutions to build a representation of the "normal" human brain.
- Fusion – experiments to control fusion are underway in multiple nations, with data shared to facilitate progress.

Each discipline is characterized by the need to access data sources (sensors, experiments, simulations) that are nationally distributed, aggregate the data into either central or distributed catalogs, and then support analysis by replicating data at multiple sites. The collections contain billions of digital objects, organized by domain specific attributes, and are accessed over the Web. To maximize utilization of the data, mirror sites and replicas are created where the research is conducted. Support for multi-lingual access is minimized by using domain specific terms and by picking a single language for describing the scientific content. Of great interest is the ability to use alternate representations for both accessing the collections (multi-lingual support) and for organizing the data (multiple ontology support). Despite having a central repository, the high-energy physics community is already faced with over 200,000 different organizations of pertinent data subsets. This constitutes a massive ontology mapping problem for deciding which organization is appropriate for a particular research question. The research issues identified in the US-Korean joint workshop will greatly benefit the SMET global digital libraries.

The need for global digital libraries for social sciences is just as critical. Global digital libraries can serve as the primary data resources for understanding cultural heritage, developing educational curricula on world cultures, and providing multi-lingual corpora to support translation research. The scale of data is also huge, when images are taken of every page of every book or report. A single text collection can be terabytes in size, with millions of images. The technologies developed to handle the scale of scientific digital libraries, can be usefully applied to social science digital libraries.

    - 5.2.2. Bioscience
  - 5.3. Education
    - 5.3.1. Working Toward a Vision: ISDL (International SMET Digital Library)

Possible Section 5.3 - RBA
While the U.S. is actively pursuing the use of information technology in the NSDL, science education is an international concern. The bi-lateral Korean-American effort can be a model for international collaboration in SMET education and a step towardan International Science Digital Library (ISDL). As with the U.S. NSDL, a joint Korean-American effort to extend the SMET educational digital libraries could be aimed at a broad ranges of users by providing rich resources and services into primary and secondary education through universities and to broader needs for scientific literacy and education by the public.

Educational content creation and access were discussed during the workshop as key dimensions for this work. Technical issues such as translation and cross-language information retrieval, ubiquitous access by wireless devices, personalization of educational resources, and the re-purposing of multimedia content would be central to meeting these two dimensions. Beyond the bilateral SMET digital library, education also provides a framework for additional research directions. Network technologies would support the collaboration of students in the two counties. In addition, education will be a major use for the Cultural Heritage resources.

- 6. Approach
  - 6.1. Initial Collaboration
  - 6.2. Program of Activity
- 7. Recommendations
  - 7.1. Near Term Opportunities
  - 7.2. Assembling Scale and Critical Mass
  - 7.3. Funding Levels and Strategies
    - 7.3.1. Planning Grants
    - 7.3.2. Balance of Small / Medium / Large Projects
    - 7.3.3. Cross-Divisional NSF Support
      - 7.3.3.1. DUE
      - 7.3.3.2. CISE
  - 7.4. Cooperation with Other Agencies' Programs
      - 7.4.1.1. DARPA
      - 7.4.1.2. NLM / NIH
      - 7.4.1.3. IMLS
      - 7.4.1.4. NEH
  - 7.5. Leveraging Existing NSF Digital Library Projects

One of the goals of the existing NSF Digital Library Phase II projects is the development of interoperable services. A particular display or query service developed within one digital library should be usable with the data collections managed by a second digital library. Three of the DLI-II projects are collaborating on the development of interoperable services, with the technology demonstrated under production load at the San Diego Supercomputer Center, and then transferred into production use for the California Digital Library (CDL). This model for technology transition can be usefully followed for the development of global digital libraries. A prototype service can be developed and tested locally, then applied within a single production digital library setting, before transference to the global digital library.

The NSF DLI-II efforts also are using production systems to identify the primary capabilities needed to create a viable global library. For the CDL effort, four critical capabilities have been identified:

- Support for persistent identifiers
- Support for replicas of data sets stored at multiple sites
- Archival storage interfaces
- Support for persistent collections

Research is conducted within the DLI-II projects to develop infrastructure to provide each of these capabilities. It is worth noting that the same critical capabilities are needed for a global digital library. Persistent identifiers are needed that are independent of the storage location and local storage protocol. For a distributed collection, this means that the name of the local file and the local access protocol need to be attributes that are stored with each data object. Support for replicas implies the ability to identify the official original version, as well as maintain consistency between the replicas. Integration with archives requires the ability to maintain a copy in an archive as well as on disk for immediate access. Finally, persistent collections must handle technology evolution. The infrastructure components needed to support these capabilities are all present within the proposed architecture for the US-Korean digital library.

The development of a multi-national digital library will identify additional key research areas, as outlined in the previous sections. The practical experience gained by building early prototypes for a multi-national digital library can be used to focus future research activities, and provide the impetus for future NSF solicitations.

7.6. Evolution to Bilateral Program

8. Conclusions