DEVELOPING A KNOWLEDGE MANAGEMENT SPIRAL FOR THE LONG-TERM PRESERVATION SYSTEM ON SEMANTIC GRID

Johannes K. Chiang[1], Eric Yen[2]

Department of Management Information Systems,
National Chengchi University, Taipei, Taiwan
[1]jkchiang@nccu.edu.tw
[2]Eric.Yen@twgrid.org

ABSTRACT

The goal of Long-term preservation (LTP) is to make the sustainability of archives lasting for a foreseeable enough time. The efforts are primarily hampered by challenges such as missing of standards, formal methodology and workflow model during archiving. This research is aiming to explore the LTP of various kinds of documents independently from the evolution of time and changes in techniques within digital environments. Basic requirements come from integration of storage management and information management, securing preservation of data, metadata, indexes, etc. This paper presents the evolutionary development of the LTP process for Governmental Archive Management and Knowledge Management. Effective search to resources and efficient storage/access on data, recovery drawing on co-location back-up, dynamic regulation on authentication and security management are tasks followed. Then, a pilot Semantic Data Grid and service matching mechanisms are described, where the ontologism plays a crucial role.

KEYWORDS

Knowledge Management, Long-term Preservation (LPT), Semantic Grid, Digital Archive Management, Ontology

1. INTRODUCTION

Long-term preservation (LTP) of digital objects is getting more and more compelling as we are producing huge amount of digital data every day. More importantly, the government funded projects to digitally archive government documents, cultural heritage, and research repository are demanding. The preservation of digital information is not just to save files into DVD or dump data to a file server or a data center. Preservation is to secure the value of the digital contents and make its full value sustainable. LTP is to make the sustainability with respect to knowledge management last for a foreseeable enough time. Anticipation for the long-term preservation is to achieve one hundred or two hundred years term, engaging to use of IT and infrastructure to make the preservation more robust, automate and ease to program as desired [1 - 3].

Sundarapandian et al. (Eds): ICAITA, SAI, SEAS, CDKP, CMCA, CS & IT 08,
pp. 173–184, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2515

Challenge of LTP results from the following problems but not the last[6, 9, 10]:

- Obsolescence of technology, media, data format, facility, process and knowledge etc.
- Workflow dominates the way to preserve the data rather than the technology. Thus no single technology would meet the requirements of variant types of digital archives.
- Content owner is reluctant to adapt best practices from paper-based archives, without any policy support or incentives.
- Fail to keep interoperation within organization and leads to fragmentation.
- Hard to have accurate cost-benefit analysis model to rationalize a best way for long-term preservation.
- Formal method and standard are still missing for long-term preservation.

Under such circumstances, the best way to support long-term preservation is cutting in with infrastructure approach and endeavoring to integrate services regarding workflow into the middleware. In our study, both the experiences for the government document archives and the National Digital Archive Program (NDAP) [6] in Taiwan are taken into account, on top of which production services have been delivered for the NDAP-LTP since 2003. For this purpose, DataGrid is deployed as the infrastructure for LTP to recruit underlying automated replication services under a global name space and certificate-based security framework. Both distributed storage management and discovery services were provided. Flexibility to incorporate workflow management, formal metadata elements and new technology adaptation are reserved as well.

In this paper, the basics of Grid technology, LTP workflow and semantic level enhancement are introduced in the second part. The discussion of LTP infrastructure and services design is followed. Services and system functionality was evaluated in the fourth part. Related study was presented before the summary and future works.

2. Seurvey On Ltp Workflows And Semantic Datagrid

Up-to-date, the LTP efforts are primarily hampered by: (1) no systematic approach for workflow management from archive management, content reuse and value-added process, and integration/interoperation of archives, etc. (2) no standard and formal methodology for a longer term preservation of digital objects, (3) policy heterogeneity of different organization; (4) no cost-benefits analysis model is applicable form management point of view; (5) no suitable measurement for information quality to make sure the required level of quality is archived and the loss of quality is negligible.

During the development of these missing components, the best policy to provide LTP services is to focus on the common denominators of infrastructure and the integration of workflow management mechanism based on variant archive related parties. In the following, we structure the state of-art in two parts, viz. (1) survey on the preservation workflow and (2) the deployment into Semantic GRID.

2.1. Workflow in LTP and its Administration

We have conducted survey on two archiving and preservation authorities, viz National Achieves Administration (NAA) and National Digital Archive Program (NDAP) within Taiwan. Researches into State of Art with respect to USA, UK, and Australia can be referenced in [10] and [13].

National Achieves Administration (NAA) in Taiwan is an official authority body which is responsible for governmental archives, mainly with identification, appraisal, access and disposal of official records and archives (Figure 1).

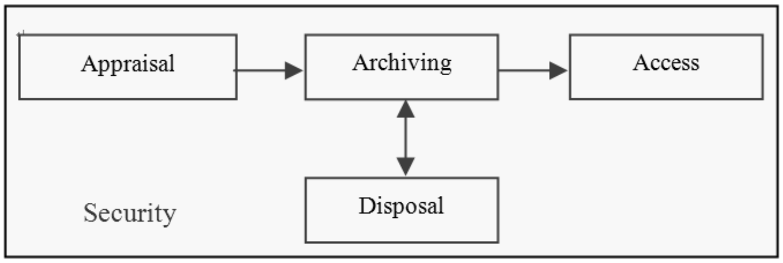

Figure 1. General Archive Administration Process by the Government[10, 14]

Since late 1990s, NAA has recognized the importance of the digital techniques with and the changes IT may bring with. “Guideline for electronic records and achieves administration” proposed by Taiwan’s National Achieves Administration (NAA), codified the achieve management process with following activities [10, 13, 14]:

1. **Ingest and Identification**: including ingest of records, contents and metadata, identifying and validation of e-signatures and timestamps, moves to storages, etc.
2. **Naming and Registration**: including naming and Registration, classification and storage, setting and registration of metadata, examination and maintenance of e-files that regards transpose, update, simulation, packaging and backup of e-records, etc.
3. **Cleaning and Appraisal**: including examination of information on records (time stamps、access rights and record sizes etc.), Appraisal and Disposal, Maintenance of Document Registries, etc.
4. **Usages of preserved Achieves**: including query of metadata, access of records, authorization of copyrights and watermarks, outputs in paper and electronic forms etc.
5. **Assessment and Secure**: including Access Right management, encryption, records of assessments, etc.

With respect to unified workflows for digital archiving, there is usually a confrontation between the counterparts of digital archive preservation, i.e. the archivist and IT professional. Archivists demand more on identification and appraisal of the records according to the contents and the retention of the originals. On the other hand, the IT professionals concentrate more on the digitization and retention of already identified objects and validate the objects in relation to the digitized forms. The typical workflow proposed by the NDAP is mainly based on the perspective of archivists as well as general user and can be summarized as follows:

1. **Accumulation:** identification, appraisal and collection of objects to be archived based on the curation policy.
2. **Material Organization and Description:** analyses on metadata, providing their annotation and interpretation, and constructing the linkage based on the collection-level metadata.
3. **Digitization:** to collect as complete as possible the digitized forms for store, manipulating and editing.
4. **Verification:** Validating the digitized objects and the annotations of good representing quality with respect to the original object.
5. **Accessing:** design of multi-accesses, including browsing and query, for users to obtain information they needs precisely.
6. **Dissemination:** providing information services and products through any means to publish the objects.

In addition to information infrastructure that centers on integrated distributed resources should be established to achieve the goals of LTP. On top of it, extensive IT supports and integration into NDAP range were concluded as follows:

- Standardization of digitization process and establishment of resources management system.
- Distributed data resources and editing organization interface
- Establishment of integrated associated index structure
- Extensible structure of distributed data resources
- Structure of integrated data storage system
- User-friendly design of browsing guide and query functions.

2.2. Deployment of the knowledge-based Data GRID

Grid infrastructure is the most viable solution for LTP at this moment, as the Grid has been proved to be production quality and sustainable by many e-Science projects in the world. Semantic Web and Ontology based information retrieve approach show the potential to offer promising means of identifying and finding target information [8, 9]. However, these potential are not exploited to their fully extent in the practical solution of aforementioned LTP problems.

The concept "Semantic Web" can be defined as an extension of the existing Web whereby information is considered with priori well-defined meaning, enabling computer and people to work in cooperation centric to Internet [6, 7]. The aim of such kind of techniques is to enhance ill-structured contents so that it may be machine interpretable or human interactive for their potential usages. On the other hand, the term "Ontology" represents a specification of conceptualization. In practical applications, ontologism provides a vocabulary for a domain and defines the meaning of the terms and relationships between them [7]. In this research, ontology is referred to as the shared understanding of domains of interest under certain sense of management, which is often conceived as a set of concepts, relations, axioms etc. The concepts within the ontology are organized in taxonomies. Figure 2 illustrates the overall architecture of the semantic GRID we employed, i.e. a hybrid of semantic web and GRID services.

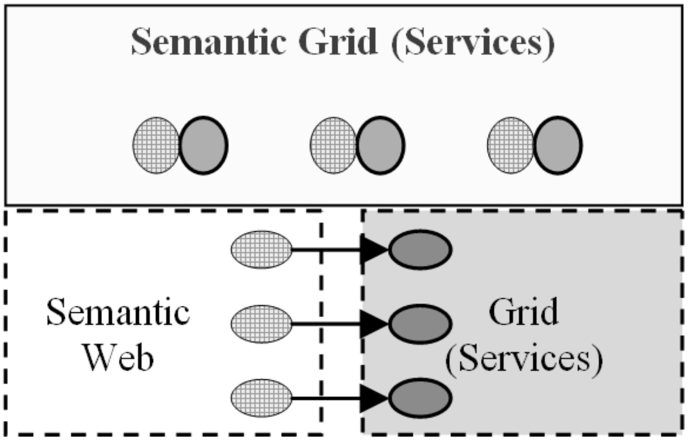

Figure 2: Overview of Semantic GRID Concept

3. REALIZATION OF THE SEMANTIC DATAGRID FOR LTP

Since late 1990s, we started to joint the NDAP and take parts in the project for LTP and Authorized Procedures for of NAA around 2000. Related development process last several years and evolutes in four phases, viz.

[1] Phase 1: Digitalizing and reformatting of Archives
[2] Phase 2: Text- and Meta-Searching on Intranet/Web
[3] Phase 3. SRB/SRM, Knowledge-based DataGrid
[4] Phase 4: Long-term archive on Semantic GRID

Within the evolutionary development, the basic concept perceived is to develop the LTP towards Organizational Knowledge Management (KM). Based on this concept, the research and development can be structured in conceptual Model, Logic Model, Operation Model and KM services. Among others, a crucial element regarding the work in the 3rd phase is the so-called "three-copy" strategy which was initiated by R. Moore [15 - 17] and extended in this research [9, 10]. These are to be clarified in the following context.

3.1. The conceptual model of LTP centric to KM

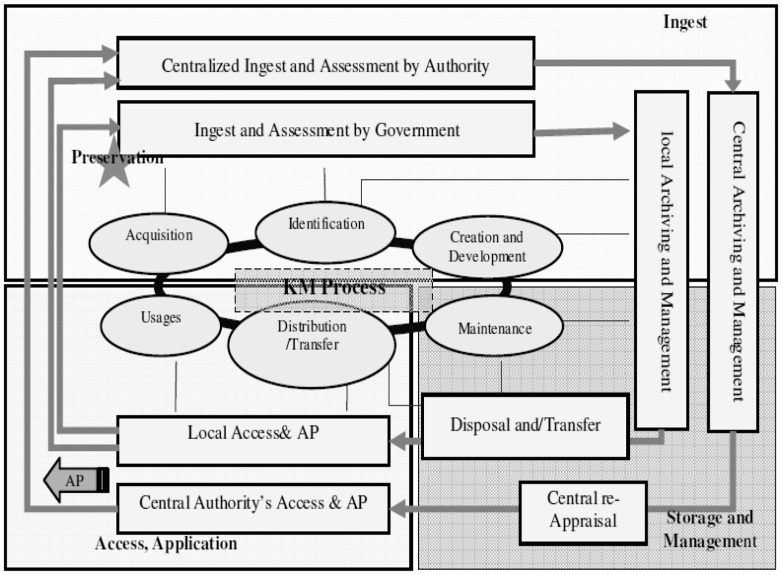

Figure 3: Knowledge Management Spiral for Achieve Preservation

To achieve the LTP centric to knowledge Management, we couple the workflow of Government Achieve Administration and NDAP with building blocks for Knowledge Management proposed by Probst et al [11]. With respect to different authorities by various government agencies, the knowledge centric achieve management system consists of two loops, that forms a spiral (Figure 3). The inner loop stands for the workflows by normal government agencies and local governments. The central authority such as the National Achieves Administration should be responsible for the activities of the outer loop.

3.2. Logical architecture for the LTP DataGRID

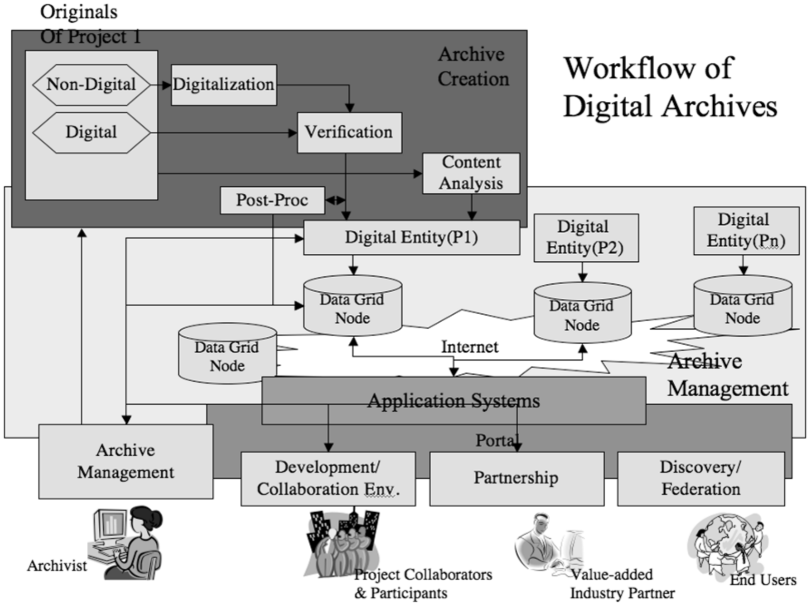

Figure 4: Logical Architecture of the NDAP LTP DataGrid System

The LTP as well as related knowledge services as a whole, with respect to aforementioned conceptual model, is realized on the top of the semantic grid. The Semantic DataGrid system we constructed enables users to create, manage, and collaborate with flexible, unified "virtual data collections" that may be stored on heterogeneous data resources distributed across a network. Logic architecture of the NDAP LTP DataGrid system could be found at Figure 4.

Basically, OAIS model [4, 6] is a reference behind the scene other than the workflow requirements from the archivists, although OAIS is just a conceptual model for digital objects preservations.

As we mentioned at first, flexibility to adopt new system architecture and new technology while meeting the workflow requirements would be the key for a LTP system. A new approach to keep track and produce the evolving meaning of the digital objects should be included as a core service. Relationship between digital objects other than the inherit collection association should be allowed to extend by the users dynamically. Thus, innovative correlations among digital contents would be discovered and new value would be generated as well. This approach introduced the Web 2.0 concepts to make the added value of digital archives with enhanced synergy.

Success of long-term archive has to take care of the whole process from appraisal, digitization process, content analysis, to the representation, user service and migration plan. Taking advantage of Data Grid technology, archival quality, reduction of operational cost and long-term access to the greatest extent value of the original collections could be further enforced. Based on the reliable LTP production system, we have constructed, schema integration, provenance trail, spatial and temporal encoding, rights management, and context-based parsing and inferences are the current focus for semantic level information services and value creation.

In our implementation, semantic level annotations are kept based on the dynamic association from real use cases, the ingestion of contents by the tracking of evolving values and selforganizations. All these newly generated interpretation and relations are recorded dynamically into the object metadata described by RDF at this moment. All the semantic contents could be associated directly to the global file catalog. User could search all the contents based on basic static field or keyword data, or get the derived targets based on the relationships.

3.3. The three-copy strategy as an underlying mechanism

As the Grid has been proved to be production quality and sustainable by many e-Science projects in the world, Grid infrastructure is the most viable solution for LTP. In terms of LTP, the off-the-shelf facility with open and reliable operation system and standard TCP/IP protocol are necessary for the base layer of the infrastructure. Upon that, the DataGrid middleware would support automatic replication under a global name space and global metadata catalog and unified access. Thus, three-copy strategy, as shown in Figure 5, is implemented with standardized metadata annotation in collection and record level respectively. Distributed storage resources management and certificate-based security model are also provided on-the-fly by the DataGrid itself. Workflow and data discovery services would be supported right above the middleware as well. DataGrid here is for the integration of Internet resources and to collaborate by means of virtual organization (VO). Semantic Grid services are integrated to enable effective discovery, automation, integration and reuse of information across applications upon Grid by giving information well-defined meaning.

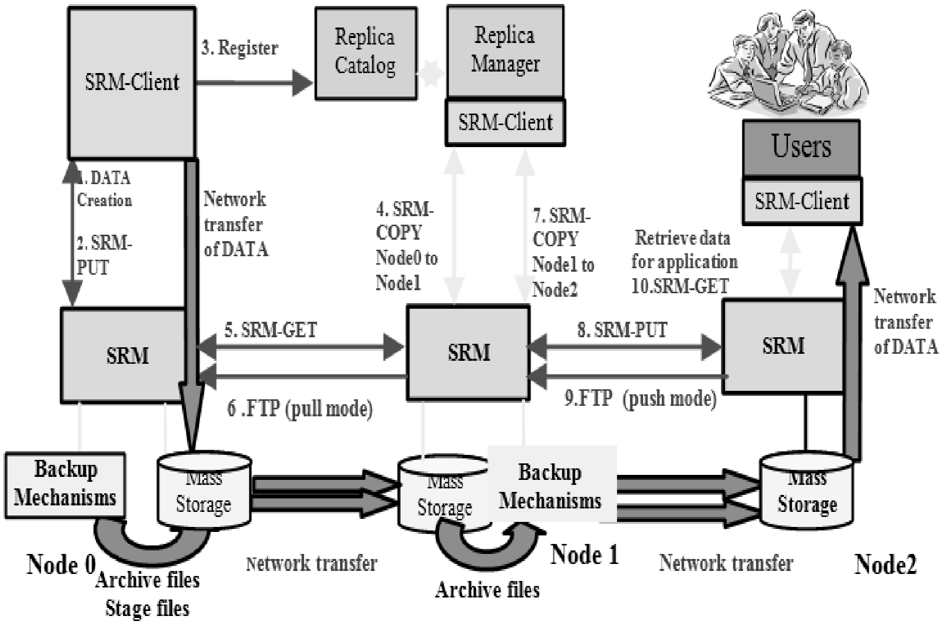

Figure 5: Three-copy strategy for Federation/ Collaboration/Synchronization

3.4. The operation model for the LTP

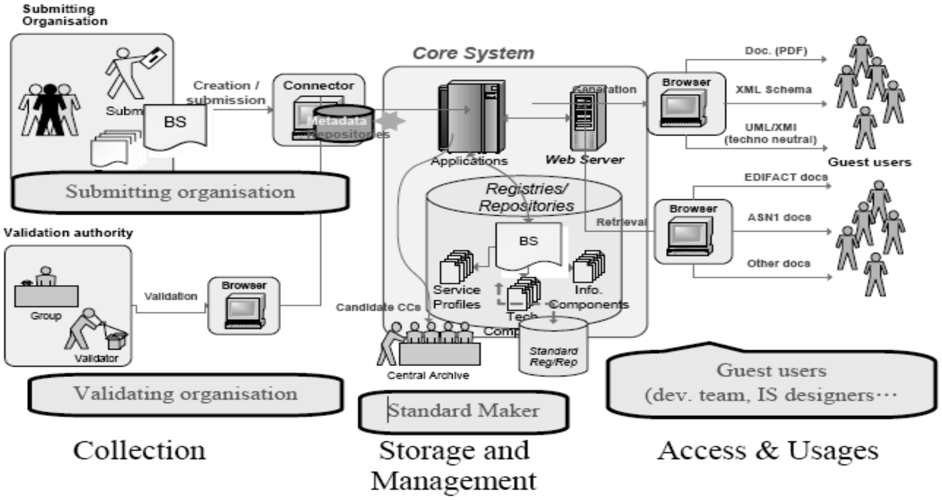

Figure 6：Knowledge-based Document Process Operation Model

Drawn on Zachman Framework [5, 9, 10], a comprehensive workflow model with respect to the government document management spiral (Figure 3) was devised as the Figure 6. In order to deploy the operations corresponding to semantic GRID, a role-based analysis for the perspectives of 5W1H with respect to the architect were implemented as well. Accordingly, a blueprint was formulated, which covers the responsibility of planner, owner, designer, builder and sub-contractor, and the correspondence of organization, activity and tasks would be defined (ref. Appendix Table 1). Abstractions of those tasks were further exploited to have clear coverage of required functionalities. Detailed services can be developed based on the defined jobs of each role at each step. Flexibility to accommodate new standards and implementation details are all preserved and be made transparent to all the parties involved, once the service interfaces are confirmed. Furthermore, the workflow could be re-compiled whenever necessary. Details of the knowledge-based collaborative document process role model could be found in the Table of the Appendix.

3.5. Implementation of Ontology and Semantic Grid

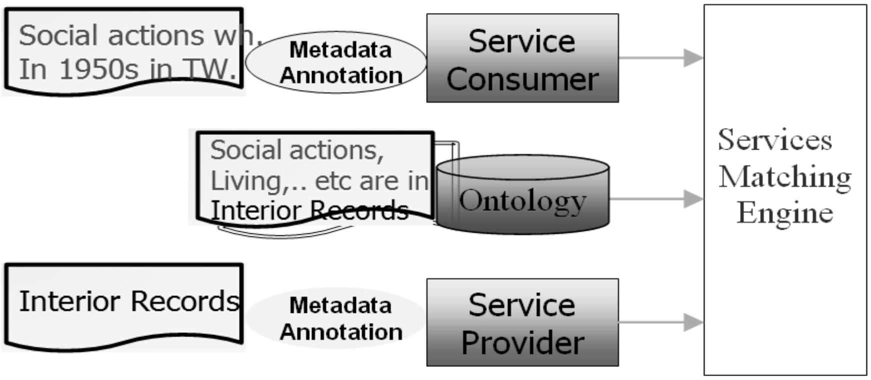

Figure 7: Semantic GRID and Semantic Service Matching

Based on the design as described above, a prototype, as shown in Figure 7, was implemented by deploying semantic web over grid as well. A simplified ontology management mechanism was also implemented, with the example of a set of ambiguous vocabulary in Chinese. Each term of this vocabulary set would be conceived as more than to different meanings. Based on the ontology composed of Dublin Core metadata and relationship property, the correct meaning could be identified. Correlation between them and the relationship are also clearly annotated and could be derived. In our pilot framework, an ontology class description is appended to every searched result to support the semantic level annotation for item. And, we use a multidimensional information/knowledge discovery algorithm as the core of the Service Matching Engine as described in [9, 18].

Figure 7 shows an example, in which a service consumer is querying information about social actions in 1950 in Taiwan. The information required can be annotated with metadata in any form/schema within the repository to describe the service asked. A knowledge discovery service provided by a government agency on the other hand can be annotated with metadata describing the service. While the Services Matching Engine receiving the two metadata sets, the engine accesses the ontology which clarifies the “social action” is a part of interior records in the context. With the result of matching, the engine will make an inference whether the discovery service can satisfy the query service request.

3.6 Security and Authentication Proxy

Nevertheless, there left still the security issues as the most service oriented developments should notice. We have proposed the Archive GRID Security Infrastructure as in Figure 8 illustrates, which is an extension of PKI on the GSN (Government Service Network) developed within the e-Government Project in Taiwan.

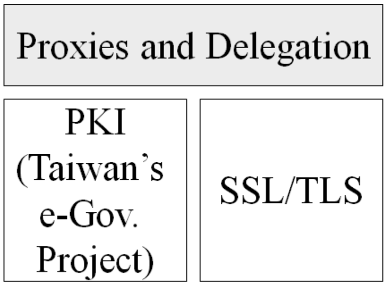

Figure 8: The Archive GRID Security Infrastructure

Upon the infrastructure, the Authentication for Resource Sharing can be defined for different service consumers. For the e-Government System, we recommend to define the notion for User Right in terms of “contents + Services”, and to delegate the right to six levels of user roles as follow (also ref. Figure. 6):

1. System Administrators
2. Validation Authorities
3. Schema Generators
4. Document Generators
5. Official Reader
6. Reader in Public (Citizen)

4. SUMMARY AND FUTURE WORKS

Long-term preservation is a totally new field for research. Although there are challenges for LTP at this moment and makes the policy, technology, infrastructure, and methodology varies a lot. The philosophy we have is to think from the infrastructure and workflow points of view. The basic concept we perceived is to give the meaning with respect to knowledge management. A semantic DataGrid system is constructed to support remote backups with the three-copy strategy and fulfill longer term preservation by migration approach. Requirements are based on the government document archival by NAA and the National Digital Archive Program in Taiwan. Based on the standard procedure to annotate digital object by standard metadata scheme, fruitful basic attributes of each data object are recorded. Web 2.0 approach is applied to keep track of the relationships between objects based on real user cases. Semantic level information would be generated and recorded for more innovative knowledge. At this moment, efficiency is not the issue to be taken into account for LTP. The theme of the research lies in the scalability and effectiveness for the tolerance and migration of new technologies into the LTP. The concept and gestalt we presented in this paper becomes one of the most significant developments for LTP in the world.

Information quality is still the most essential issue for a digital archive system in terms of the user point of view. Objective measurement of information quality could help evaluate the collection and digitization/annotation process to see if the information quality is good enough from the beginning. Similarly, information loss through time or during transformations should be measured to avoid information decay before realized. Formal model for a cost-benefit analysis of digital archive is still very useful but not applicable. More flexible semantic level information integration between archives and with other resources should be improved in the future.

5. ACKNOWLEDGEMENTS

This work was partially supported by National Science Council Project for e-Science Asia Centre and Federation in Taiwan (NSC 94-3111-P-001-002-Y02).

REFERENCES

[1] Supplement to the President's Budget for Fiscal Year 2007. The Networking and Information Technology Research and Development Program, A report by the Subcommittee on Networking and Information Technology Research and Development, Committee on Technology National Science and Technology Council, February 2006.
http://www.nitrd.gov/pubs/2007supplement/07SuppEntireBook/07Supp_FINAL-022306.pdf

[2] The National Digital Information Infrastructure and Preservation Program (NDIIP), 2006. http://www.digitalpreservation.gov/

[3] National Archives and Records Administration (NARA) usa. http://www.archives.gov/index.html . 2006.

[4] ISO Archiving Standard - Reference Model for an Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0b1.pdf

[5] W. H. Inmon, J. A. Zachman and J. G. Geiger, “Data Stores, Data Warehousing, and the Zachman Framework: Managing Enterprise Knowledge”. Mcgraw-Hill(1997).

[6] National Digital Archive Program, http://www.ndap.org.tw/

[7] Jena – A Semantic Web Framework for Java, http://jena.sourceforge.net/.

[8] Maozhen Li, Mark Baker: “The GRID – Core Technologies”, Willy 2005.

[9] Johannes K. Chiang: “Developing a Governmental Long-term Archive Management System on Semantic Grid”, PNC Annual Conference and Joint Meetings, Oct. 2005, Hawaii USA.

[10] Johannes K. Chiang, Simon C. Lin, Eric Yen, Yin-Ru Lai, “Investigation of for the Architecture and technological procedures for long-term preservation of digital archives”. Research Report, National Archive and Records Bureau, ROC, Taipei, 15 Dec. 2004.

[11] G Probst, S Raub, K Romhardt, “Managing Knowledge -: Building Blocks for Success”, ISBN: 0-471-99768-4, 2000.

[12] Thibodeau, K., “Building the Archives of the Future”, D-Lib Magazine, 2000.

[13] Pei-Ing Chao, “A Research into Governmental Archive Administration in Australia, UK, USA and Taiwan”, Master Thesis, EMBA National Chengchi University, 2003.

[14] Yin-Ru Lai, “Collaborations in knowledge-centric Archive Management”, Master Thesis, National Chengchi University, 2005.

[15] R. Moore, ”Using Data Grids to Manage Distributed Data”, Lecture slides for PNC 2004 Annual Conference, Taipei, Taiwan, 17-22 October, 2004.

[16] R. Moore and A. Merzky, “Persistent Archive Concepts”, Global Grid Forum Informational Memo GWD-I, Sept, 2003.

[17] R. Moore, “Knowledge-based Grids”, Proceedings of the 18th IEEE Symposium on Mass Storage Systems and Ninth Goddard Conference on Mass Storage Systems and Technologies, San Diego, California, April 2001.

[18] Johannes K. Chiang, Gestalt of an Approach for Multidimensional Data Mining on Concept Taxonomy Forest to Discover Association Patterns with various Data Granularities, Proceedingsof the 19th IEEE ICTAI 2007, Patras Greece, Oct. 29-31, 2007 (to be published)

APPENDIX

TABLE : The Role Model for Knowledge-based Collaborative Document Process

	Component	Activity	Place	
	What	How	Where	
Planner	Government Archives Model and digital objects	● Accumulation ● Archive management ● Access & Application	● National Archives ● Gov. Agencies ● Curation Organization	Scope
Owner	● Permanent Files ● Transition criteria ● Durable Files ● Volatile Files	● Appraisal and Classification ● Composition and Custody ● Data Cleaning ● Access and Application	● Checkpoint ● Backup point ● Deep archive point ● Management/Policy level point ● Supporting service point	Enterprise Model
Designer	● Core Schema ● Document of workflow and Specification ● Metadata - Structure - Description - Operation	● Appraisal & Classification ● Cataloging ● Curation ● Clearance & Transition ● Use Case Analysis ● Service Level Identification	● Core service ● Archive service ● Long-term preservation Service ● Deep archive ● User service	System Model

Builder	● File ● Message/document Exchange protocol ● Privilege/Role management ● Clearinghouse services	● Operation model ● Workflow of government document archive management ● Use Case implementation	Arrangement and integration of the components/Services	Technical Model
Sub-contractor	● Document format ● Data type ● Encoding system ● Message packet structure	● Metadata management ● Migration of document ● Core fields validation ● Arrangement of physical space ● Storage resource management ● Message exchange ● Operation performance monitoring ● Assurance of functionality and quality of information flow and services	System architecture (topology of core components and real-life information), workflow and services	Components

AUTHORS

Prof. Dr.-Ing. Johannes K. Chiang is now a faculty member on the Department of MIS and Deputy Director of the Center for Cloud Computing at National Chengchi University Taipei. He received the academic degree of Doctor in Engineering Science (Dr.-Ing., Summa Cum laude) from the RWTH University of Aachen, Germany. His current research interests include "Data and Semantic GRID", "Business Intelligence and Data Mining", e-Business and ebXML, Business Data Communication. He also serves as a consultant for the government agencies in Taiwan and as an active member of various international affiliations, such as IEEE, ACM etc. before 1995, he has been a research fellow at RWTH of Aachen and Manager of EU/CEC projects.

Eric Yen is a senior manager of Academia Sinica Grid Computing (ASGC) Centre in charge of grid technology, and also a Ph.D student at Department of MIS of National Chengchi University (NCCU) Taipei, Taiwan. He received the master of science degree on Computer and Information Engineering in Tamkang University, Taiwan in 1989. He works on Grid/Cloud technology at ASGC since 2002, primarily for the WLCG Tier-1 center of Taiwan, the EGEE Asia Federation in Taiwan, and also the e-Science infrastructure extension and application development in both Taiwan and Asia Pacific Region.