Communication and Storage of Data for Personal Libraries*

*Personal libraries are proposed by investigating design elements that should be considered in the creation of a **distributed digital object store** for a personal library system. Particular emphasis is placed in **storage scalability** and **communication** demands.*

Francisco Álvarez Cavazos

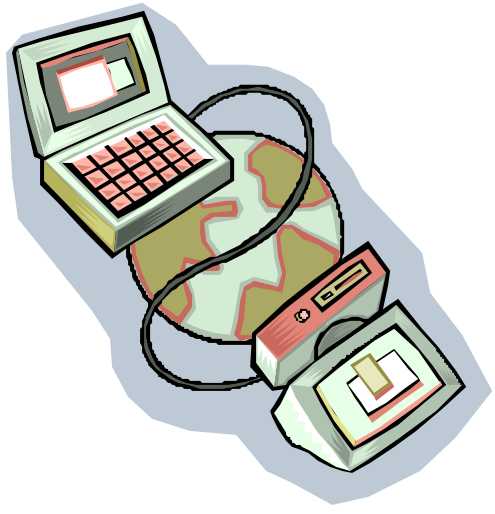
JCDL'05 Doctoral Consortium

June 7, 2005

❶Introduction

Personal Libraries: Concept and Challenges

- Vannevar Bush (1945) envisioned a great record of human knowledge
- Bush proposed a collective **“common record”** and **“personal records”** that resided in personal information machines called *Memex*
- Modern *Memex*-like technologies:
  - Personal Computers
  - Digitalization (text, speech, images, and others)
  - Internet
  - World Wide Web (WWW)
  - Digital Libraries



❶Introduction

Personal Libraries: Concept and Challenges

- However, computers dispute the *Memex* role with books, notebooks, sketch pads, adhesive notes and the rest of the paper and ink world!

❶Introduction

Personal Libraries: Concept and Challenges

- While the WWW and distributed★ digital library systems resemble Bush's common record, we have not achieved a uniform concept of personal records
- Modern-day approaches to personal records:
  - the hard drive of my PC?
  - my personal account in a networked file system?
  - a "only client" digital library installation? [e.g. Greenstone, UpLib]

❶Introduction

Personal Libraries: Concept and Challenges

Challenges of personal library technology:

I. To provide the individual user with a general purpose repository★

II. To provide selection of personal library content★ based on:
   a) information-retrieval techniques and
   b) user-defined classification and indexing schemes

III. To provide the user with the capability of accessing her/his personal digital library from anyplace at anytime

IV. To assure long-term archival of library content

V. To grant an interminable storage capacity★ for library content

VI. To allow each user to share the content★ of her/his personal library with other users

VII. To provide interoperability with collective digital library systems and/or the World Wide Web

⇒ ***Effectively a new breed of digital library systems (Borgman, 2003)***

❷Background

Personal Libraries and Digital Object Stores

- A *digital library* can be conceptualized as a collection of services built around digital objects
- *Digital library services* traditionally include:
  - document submission
  - full-text and metadata indexing
  - document search and retrieval
- Libraries are composed of *digital objects* (e.g. documents, technical reports, movies) of several media types (e.g. text, audio, images, video)

❷Background

Personal Libraries and Digital Object Stores

- A digital object has one or more binary representations (e.g. formats) and has associated metadata
- The objects of a digital library reside in a *digital object store*
- Digital object stores can be implemented atop (Mather, 2001):
  - file systems
  - databases with large data objects★
  - digital object repositories

❷Background

Personal Libraries and Digital Object Stores

- Digital libraries require specialized information-retrieval (IR) beyond the retrieval of typed data provided by database technology (Adam et al., 1996)
  - e.g. full-text search
- ⇒ To support full-text search in our digital object store proposal, a *loosely coupled integration* (Raghavan and Garcia-Molina, 2001) of database and text retrieval systems is being followed

❷Background

Personal Libraries and Digital Object Stores

- With mobile technologies and universal access, library systems with tens of millions of users will appear ⇒ architectural changes to:
  - **Communication Services** are critical since:
    - Information Infrastructure → Inherently Distributed
      - Universal Access (III challenge)
      - Content Sharing (VI challenge)
      - Interoperability (VII challenge)
  - **Storage Scalability** is demanded by:
    - +Mobile users, ++Distributed Storage Resources, +++Digital Objects
      - Individual Repository (I challenge)
      - Selection (II challenge)
      - Lifetime Storage (IV challenge)
      - Unlimited Storage (V challenge)

❸State of the Art
Personal Libraries Approaches

- Personal libraries can be categorized according to:

1. Their *personalization* support over their *distribution architecture*★ and

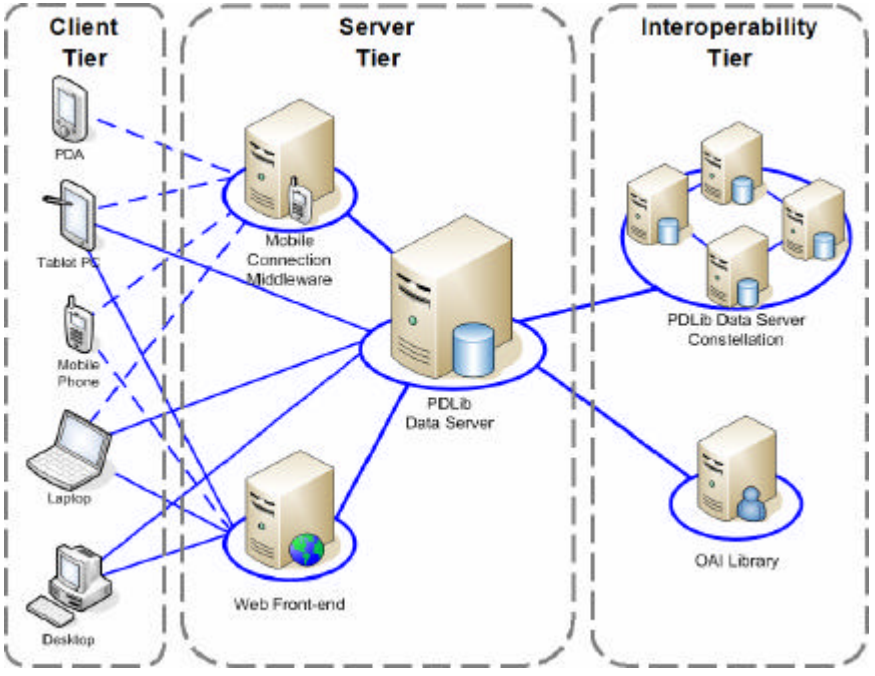
2. Their approach to library *content selection* (Challenge II)

❸State of the Art
Personal Libraries Approaches

- According to their *personalization support* over their *distribution architecture*, personal library systems can be classified as:
  - **Client-side Personal Libraries**. Provide personalization services to the user's interface with the library system (e.g. Daffodil, Paddle, MiBiblio, PoPs, MyLibrary, Salticus)
  - **Peer-to-peer Personal Libraries**. Personal digital library services are provided by the client/server nodes of a distributed peer-to-peer network (e.g. Ding's Semantic Search P2P DL, Guan and Zhang (2004)'s PDL)
  - **Middleware Personal Libraries**. The actual content of the personal libraries comes from *heterogeneous* data sources registered at a mediation environment and mapped to a user-defined schema (e.g. Briukhov et al. (2001)'s Synthesis)
  - **Server-side Personal Libraries**. Personal libraries are provided by the server infrastructure of the digital library system. The server-side approach is specially suited to support access from multiple device types (e.g. Slater's Whisper, He and Shen (2005)'s MyPDL and ***PDLib***, *testbed of this research*)

❸State of the Art
Personal Libraries Approaches

- The main component of PDLib's server infrastructure is the ***data server***★:
  - services for storage, classification and information retrieval of personal library content
  - interaction with other digital library systems via the OAI-PMH
- Other components of PDLib are:
  - **Client-side Applications**. Target mobile and fixed devices adapting digital library services to client device capabilities
  - **Mobile Connection Middleware**. Mediates the interaction of mobile devices with the data server
  - **Web Front-end**. Transforms personal library services into a Web application for browsers and microbrowsers.



❸State of the Art
Personal Libraries Approaches

- Current approaches have focused on the ***client-side***
- Libraries provided by the both the ***client-side*** and the ***middleware*** approach are at odds with the individual repository requirement of personal libraries (Challenge I)
- ***Peer-to-peer*** personal libraries are at odds with the universal access requirement of personal libraries (Challenge III)

❸State of the Art
Personal Libraries Approaches

- Mobile devices place the ***server-side*** approach as a promising alternative to provide universal access (Challenge III) since it facilitates access from multiple device types

- In addition, a digital object store can be readily designed to provide an individual repository to each system user

❹Research Question

- First, it is necessary to specify which IR mechanisms and classification and indexing schemes are considered in this research:
  - **IR Mechanisms**. We propose a distributed IR strategy that combines:
    - *full-text queries* issued to text retrieval systems, with
    - *typed queries* issued to database systems
  - **Classification and Indexing Schema**. This research proposes a *flexible collection and metadata management* (Alvarez-Cavazos et al. 2005)
    - Personal libraries are composed of collections
    - Collections contain, in turn, other collections and/or documents
    - Users interact with personal digital libraries by
      - creating and deleting collections
      - submitting, moving, copying or downloading documents
    - Users can define the metadata set that will be used in each collection

❹Research Question

- Which design elements should be considered by a ***distributed digital object store***
  - *text retrieval systems loosely coupled with*
  - *distributed databases with large objects*
- to satisfy:
  - *the **storage scalability**—required by:*
    - individual repositories (I) with:
      - flexible collection and metadata management (II)
      - full-text and typed selection of content (II) and
      - unlimited storage capacity (V)—*and*
  - *the **communication services**—required by*
    - library content sharing (VI)
- of server-side personal libraries?

❹Research Question
Who is Interested and Why?

- Research in cognitive psychology and information science has shown that individuals (Goh and Kacmar, 1995):
  - Organize, search and retrieve information differently, and
  - Tend to remember where information is located if they are the organizers
- Þ ***We all are interested in personal libraries!***

❺Preliminary Results

- The preliminary results of this research are:
  - ① The design of a centralized digital object store for server-side personal libraries and its prototype implementation in the context of our testbed: the data server of the PDLib system
  - ② The design of the loosely coupled integration architecture of text retrieval and database systems for the storage of personal library digital objects according to our proposed IR strategy

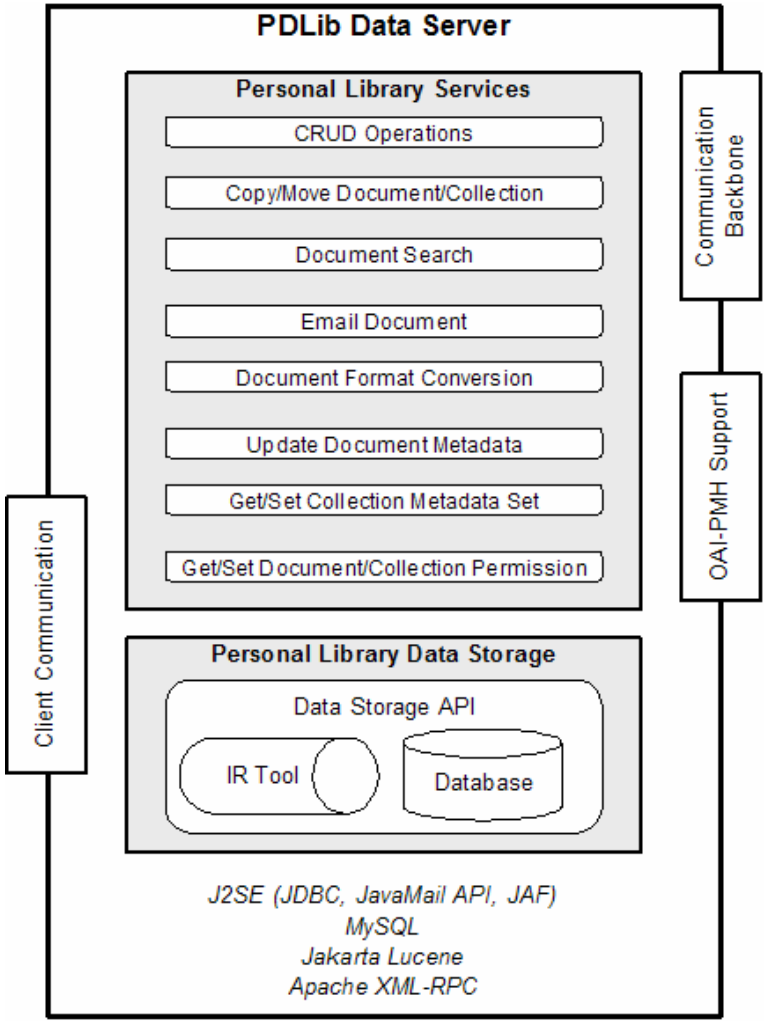
❺Preliminary Results

- ③ A design strategy to address the storage scalability and communication requirements of server-side personal libraries based on:
  - **Data compression** (). To reduce the storage demand of a digital library's data objects
  - **Data integration** (). Will harness storage resources of distributed data sources to address storage scalability
  - **Data communication** (). Provide a communication backbone in order to satisfy the library content sharing

❺Preliminary Results

① Centralized Data Server: Design and Implementation

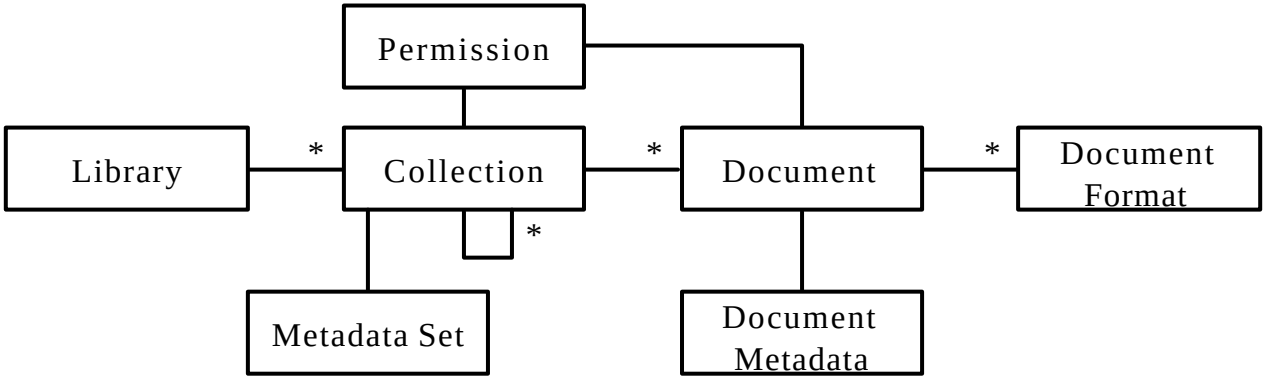
- The data server is the digital object store of the server-side personal libraries provided by the PDLib system
- The data server provides personal library services, stores personal library data and supports interoperability via the OAI-PMH[1]



[1] OAI-PMH support was developed in the context of an associated master thesis (Hurtado-Alvarado, 2005)

❺Preliminary Results

① Centralized Data Server: Design and Implementation



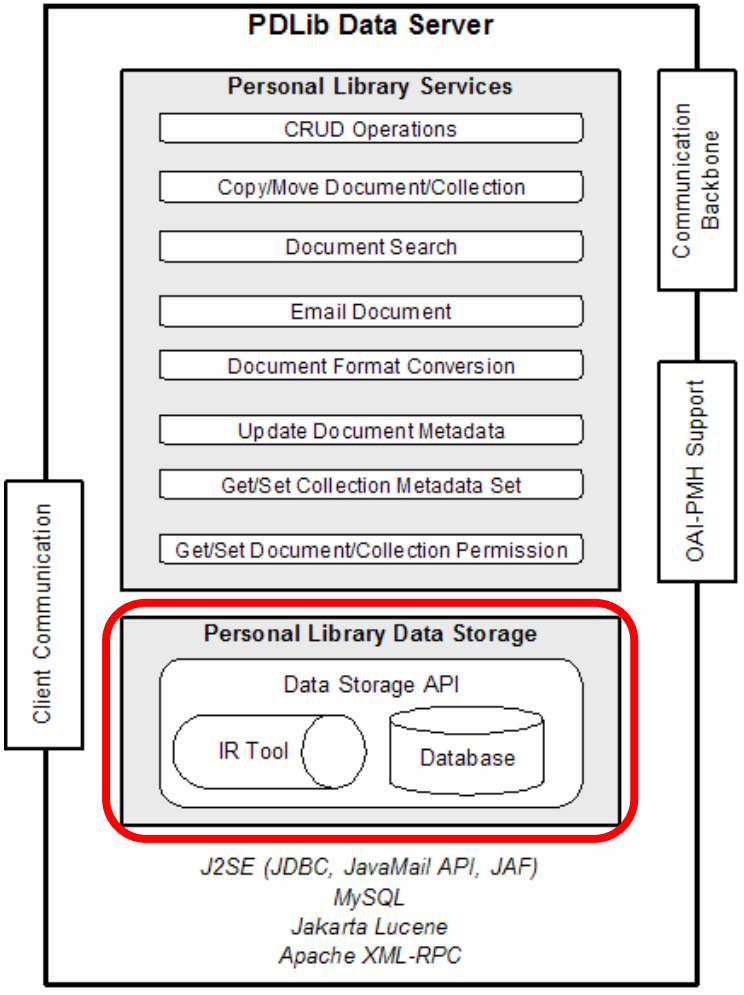
Personal Digital Library Data Model:

- A library contains one or more collections
- A collection contains documents or more collections and is associated with a metadata set
- Metadata sets are composed by one or more metadata definitions
- Document metadata is the metadata of a particular document
- Both collections and documents have associated permissions
- A document can have one or more document formats

❺Preliminary Results

② Data Storage Architecture: Design[2]

- An architecture to provide structured storage with scalable text information retrieval (IR) capabilities for personal digital libraries has been designed
- The data server's data storage component maps the data model (*collections, document formats and document metadata*★) to:
  - Relational Model (tables, rows, columns, data types)
  - Lucene's IR index structure (indices, documents, fields, text domain)

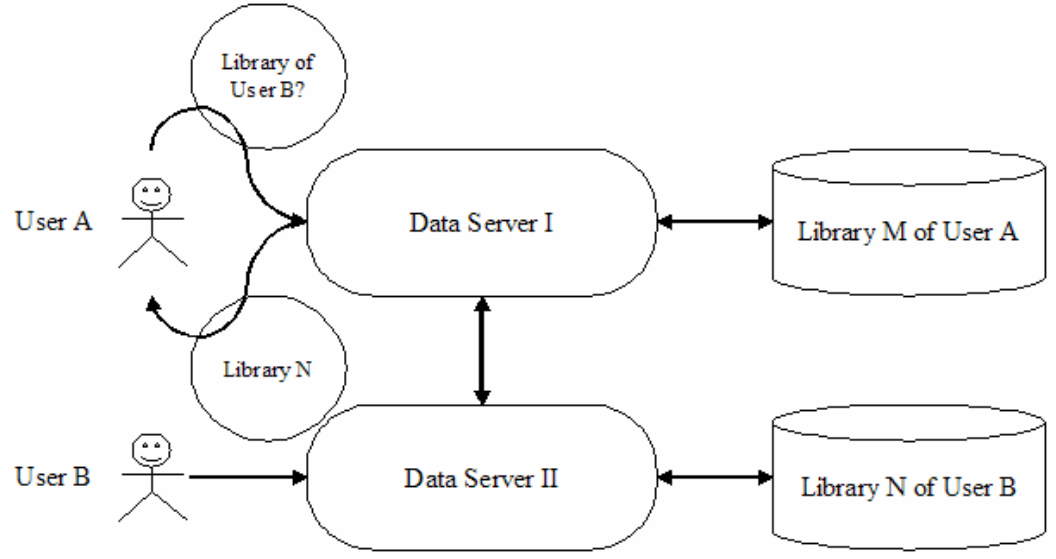


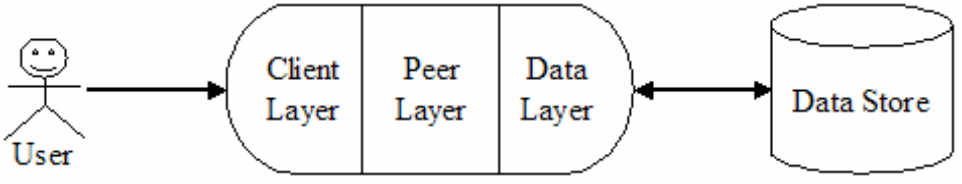
[2] Centralized implementation is underway in the context of an associated master thesis (Hurtado-Alvarado, 2005)

❺Preliminary Results

③ Data Server Future Design Strategy

- The data server is divided into design units:
  - ***Client Layer.*** Set of services accessible by a user ( )
  - ***Peer Layer.*** Communicate with other data servers if it is so required by the user's query ( )
  - ***Data Layer.*** Read and write data in the data store (centralized or distributed) ( )

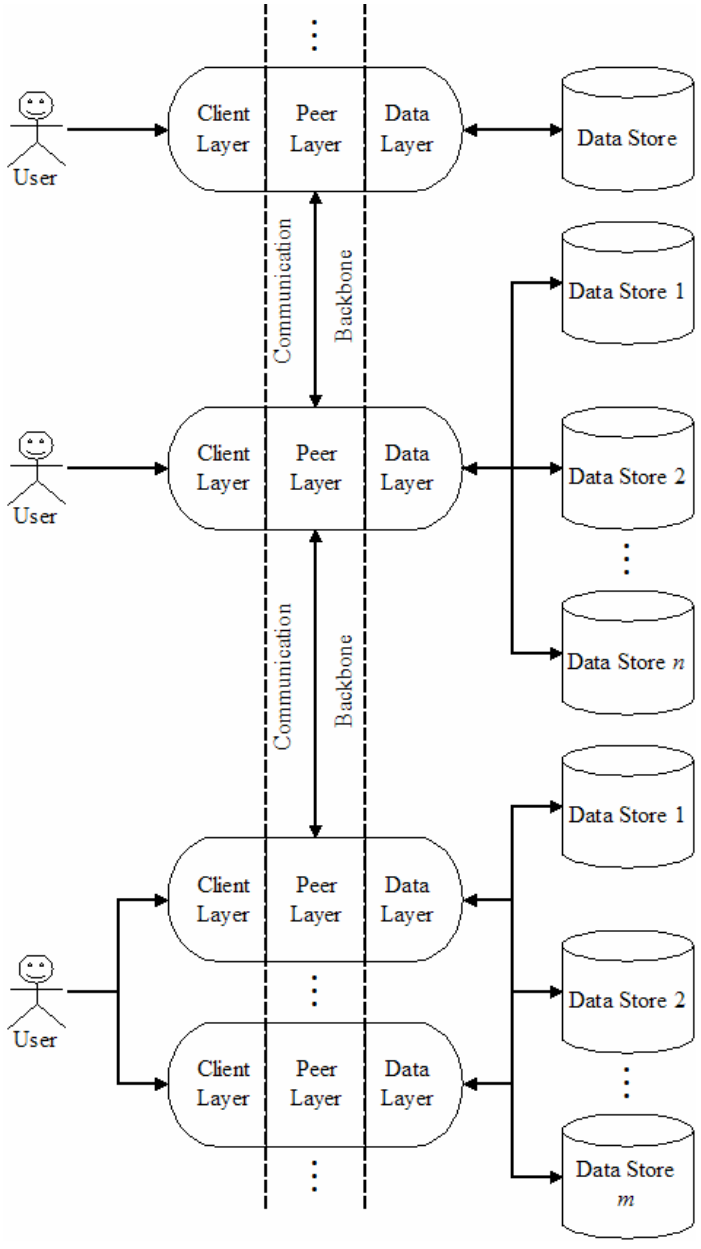




❺Preliminary Results

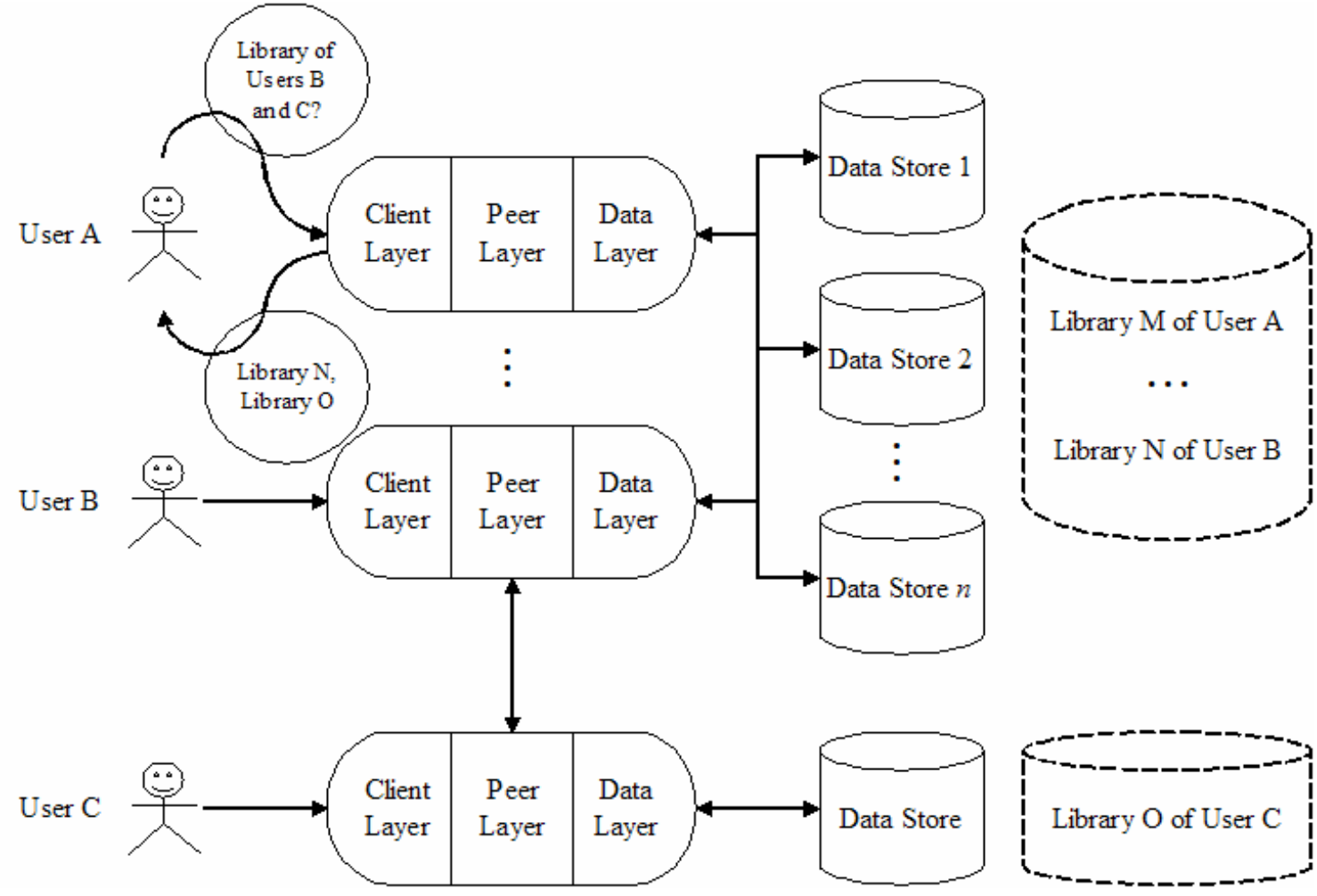
③ Data Server Future Design Strategy

- The peer layer establishes the system's communication backbone
- If two or more data servers access the same distributed database, they shall not interact directly via the peer layer since we expect higher communication performance at the data layer (, , strategy)



❺Preliminary Results

③ Future Design Strategy: The Data Server in Action!



❻Methodology

- This research follows the ***engineering paradigm*** of computer science (Wegner, 1976)

- Three-phase trial-error/success approach:

  1. Trials = ***development*** of the concept implementation of design; Error/success = agreement with functional requirements

  2. Trials = ***capacity characterization*** models; Error/success = model's accuracies

  3. Trials = ***Optimization*** of the concept implementation after capacity characterization feedback; Error/success = performance improvement

❼Expected Contributions and Technical Innovations

- The expected contributions of this research are:
  - The **reference architecture** and **concept implementation** of the
    - data compression
    - data integration and
    - data communication
  - mechanisms of a distributed digital object store, deployed over
    - text retrieval systems loosely coupled with
    - distributed databases with large objects
  - for server-side personal libraries with:
    - flexible collection and metadata management
    - full-text and typed selection of content
    - unlimited storage capacity and
    - library content sharing

❼Expected Contributions and Technical Innovations

- This research leads to the following technical innovations:

1. A scalable, distributed digital object store for server-side personal libraries

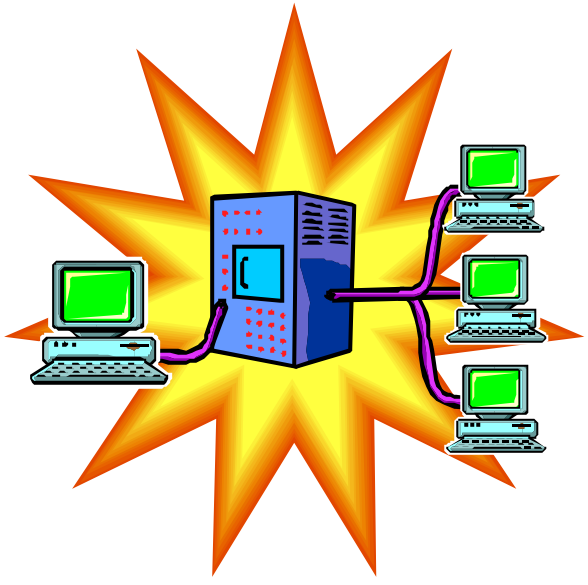
2. Personal libraries with unlimited storage capacity

3. Selection of personal library content based on:
   a) A hybrid information retrieval strategy that combines full-text queries with typed queries; and
   b) A flexible collection and metadata management classification schema

4. Personal library content sharing

❽State of Research

- Narrowing scope to ***storage scalability*** (✓, ✘, ✘)
  - ✘ communication services for library content sharing (VI)
- We are now pursuing a digital object store for server-side personal libraries (I) with:
  - flexible collection and metadata management (II)
  - full-text and typed selection of content (II) and
  - unlimited storage capacity (V)
- i.e. distributed data stores instead of distributed digital object stores



• End •

Thanks for your attention!

Comments to:

Francisco Álvarez Cavazos
Informatics Research Center
ITESM, Campus Monterrey
Monterrey, Mexico
a00782553@itesm.mx

Last updated: June 7, 2005.

❽ References

- V. Bush. As we may think. Atlantic Monthly, 176:101–108, July 1945.
- C. L. Borgman. Personal digital libraries: Creating individual spaces for innovation. Wave of the Future: NSF Post Digital Library Futures Workshop, June 2003.
- P. Mather. Scalable storage for digital libraries. Computer Science Departmental Technical Report TR-02-21, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA, October 2001.
- N. R. Adam, R. Holowczak, M. Halem, N. Lal, and Y. Yesha. Digital library task force. IEEE Computer, 29(8):89–91, August 1996.
- S. Raghavan and H. Garcia-Molina. Integrating diverse information management systems: A brief survey. IEEE Data Engineering Bulletin, 24(4):44–52, 2001.
- S.-U. Guan and X. Zhang. Design and implementation of a web-based personal digital library. Journal of the Institution of Engineers Singapore, 44(3):59–77, 2004.
- D. O. Briukhov, L. A. Kalinichenko, and N. A. Skvortsov. Personalization through specification refinement and composition. In DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, 2001.
- F. Alvarez-Cavazos, D. A. Garza-Salazar, J. C. Lavariega-Jarquin, L. G. Gomez-Martinez, and M. Sordia-Salinas. Universal access architecture for personal digital libraries. Technical report, ITESM, Campus Monterrey, Informatics Research Center, Mexico, January 2005.
- P. Wegner. Research paradigms in computer science. In ICSE '76: Proceedings of the 2nd International Conference on Software Engineering, pages 322–330. IEEE Computer Society Press, 1976.
- D. A. Menasce and V. A. F. Almeida. Capacity Planning for Web Services: Metrics, Models, and Methods. Prentice Hall PTR, second edition, September 2001.
- L. A. Hurtado-Alvarado. Interoperability and text retrieval in a personal library server. Master's thesis, ITESM, Campus Monterrey, May 2005.
- David M. Levy and Catherine C. Marshall. Going digital: a look at assumptions underlying digital libraries. Commun. ACM. 38(4):77-84, 1995.
- Crespo, A.; Garcia-Molina, H.. Archival Storage for Digital Libraries, Third ACM Conference on Digital Libraries. Pittsburgh, PA, USA, June 23-26, 1998.
- D. Andresen, T. Yang, O. Egecioglu, O.H. Ibarra and T.R. Smith (1996). Scalability Issues for High Performance Digital Libraries on the World Wide Web. Advances in Digital Libraries: 139-148. http://citeseer.nj.nec.com/andresen96scalability.html.
- Sriram Raghavan and Hector Garcia-Molina. Integrating Diverse Information Management Systems: A Brief Survey. IEEE Data Engineering Bulletin. 24(4):44-52, 2001.
- B. Bhargava, S. Li, and J. Huai. Building high performance communication services for digital libraries. In Forum on Research and Technology Advances in Digital Libraries, McLean, Virginia, May 1995.
- Bharat K. Bhargava and Melliyal Annamalai. Communication costs in digital library databases. In Database and Expert Systems Applications, pages 1–13, 1995.
- Dion Goh and Chuck Kacmar. Building personal digital libraries from network sources. SIGWEB News. 4(2), 1995.
- Ding, H. 2005. Semantic Search in Peer-to-Peer based Digital Libraries. In Proceedings of the First Doctoral Consortium of the ACM/IEEE-CS Joint Conference on Digital Libraries (Denver, CO, USA, June 07, 2005). JCDL '05.
- Slater, M. 2005. Whisper: A Collaborative Academic Work Environment. In Proceedings of the First Doctoral Consortium of the ACM/IEEE-CS Joint Conference on Digital Libraries (Denver, CO, USA, June 07, 2005). JCDL '05. (see also http://whisper.dforge.cse.ucsc.edu/)