

ECDL 2003

An Integrated Digital Library Server with OAI and Self-Organizing Capabilities

August 18, 2003

Hyunki Kim

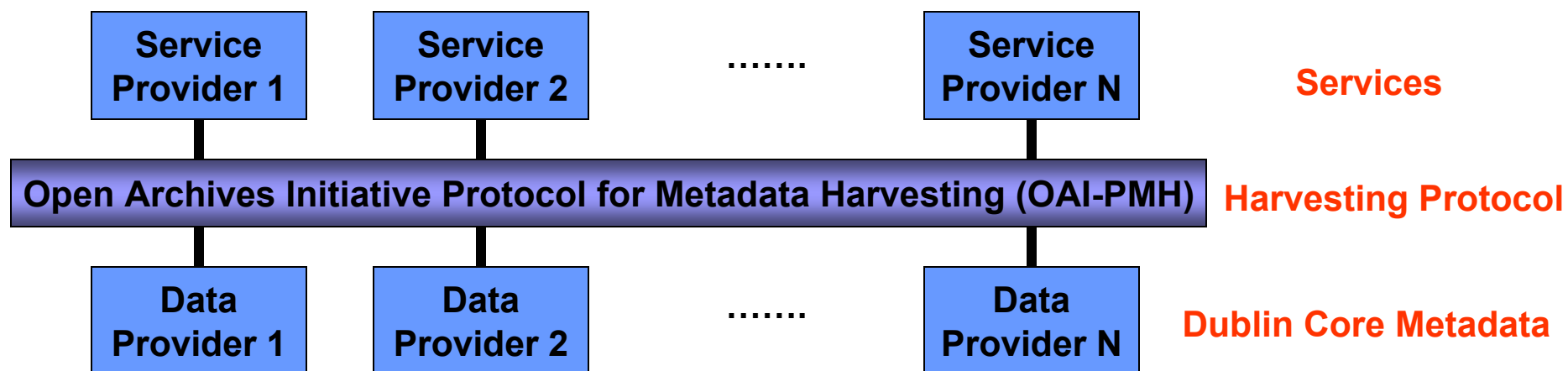


Table of Contents

- Introduction
- Related Work
- Data Mining Method
- System Overview
- Lessons Learned
- Conclusion

Introduction (1/3)

- Open Archives Initiative (OAI)
 - Experimental Initiative for the interoperability of digital libraries based on metadata harvesting
 - Collaborative effort to develop and promote interoperability standards to facilitate the efficient dissemination of digital content
- OAI Framework





Introduction (2/3)

■ Challenging Issues

- Metadata quality: XML encoding/syntax errors and vocabulary inconsistency
- Resource discovery problem with cross-archive search
 - Various kinds of data providers
 - User's inability of expressing their information needs
- Scalable architecture
- Repository synchronization



Introduction (3/3)

■ Proposed approach

- Combines cross-archive search and data mining
- Provides multiple viewpoints of harvested metadata
 - Cross-archive search provides a term view of harvested metadata.
 - Concept browsing provides a subject view of harvested metadata.



Related Work

■ Federated searching

□ Distributed information retrieval approach

- Server selection (database selection), Query processing, and Results merging
- Pros: Freshness
- Cons: Vulnerable to its weakest component, difficult apply to real world environments

□ Harvesting approach

- Distributed IR can be emulated after harvesting metadata and building a cross-archive search on the harvested metadata
- Pros: Good retrieval performance
- Cons: data duplication, repository synchronization problem



Data Mining Method using the SOM (1/5)

■ Self-Organizing Map (SOM)

- Competitive and unsupervised learning algorithm
- Artificial neural network algorithm for visualizing and interpreting complex data sets
- Providing a mapping from a high-dimensional input space to a two-dimensional output space

■ Data mining process for organizing metadata

- Document clustering using the hierarchical SOM
- Deduction of concept hierarchy
- Visualization of the concept hierarchy



Data Mining Method using the SOM (2/5)

- Data

- 19,559 metadata records from 5 OAI data providers

OAI Repository ID	Repository Name	Number of Harvested Records
caltechcstr	Caltech Computer Science Technical Reports	358
LSUETD	LSU Electronic Thesis and Dissertation Archive	324
HKUTO	Hong Kong University Theses Online	8,598
	M.I.T. Theses	6,830
VTETD	Virginia Tech Electronic Thesis and Dissertation Collection	3,449



Data Mining Method using the SOM (3/5)

- Preprocessing: feature extraction and selection

- Consider three Dublin Core Metadata elements
 - Subject, title and description elements
- Two feature vector sets for the SOM
 - Subject feature vector constructed by indexing the subject elements of the metadata collection
 - Description feature vector constructed by indexing the title and descriptor elements of the metadata collection

- Results

- Subject feature vector: 1,760 terms identified after removing 14,029 terms
- Description feature vector: 1,996 terms identified after removing 141,617 terms
- Note that 6,960 and 16,478 documents did not have the subject and description elements, respectively.



Data Mining Method using the SOM (4/5)

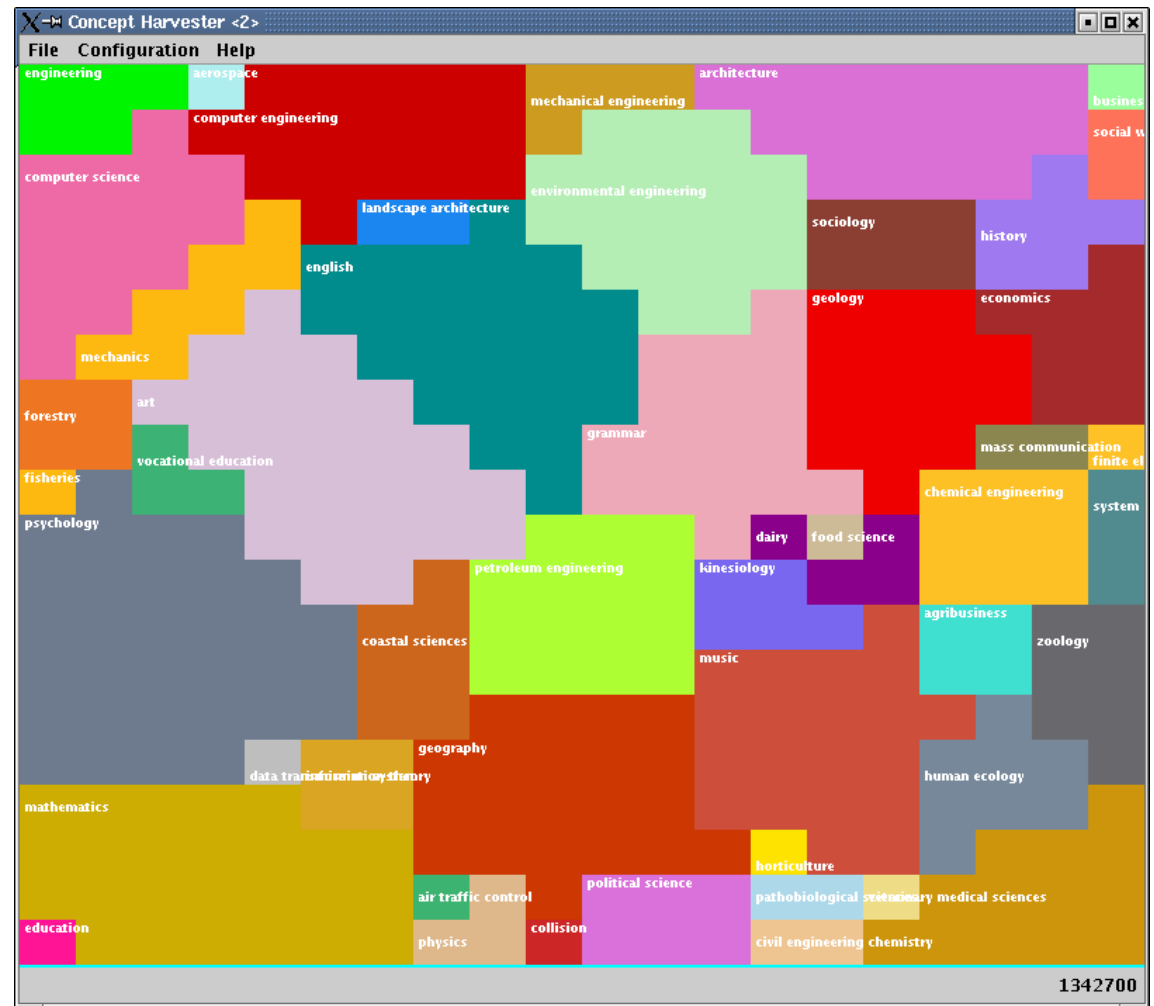
- Construction of a concept hierarchy

1. Initialize network by using the subject feature vector
2. Present input vector in sequential order
3. Find the winning node by computing the Euclidean distance for each node
4. Update weights of the winning node and its topological neighborhoods
5. Repeat steps 2-4 until all iterations have been completed
6. Label nodes of the trained network with the noun phrases of the subject feature vector
7. Repeat steps 1-6 by using the description feature vector as the input vector for each grouped concept region

Data Mining Method using the SOM (5/5)

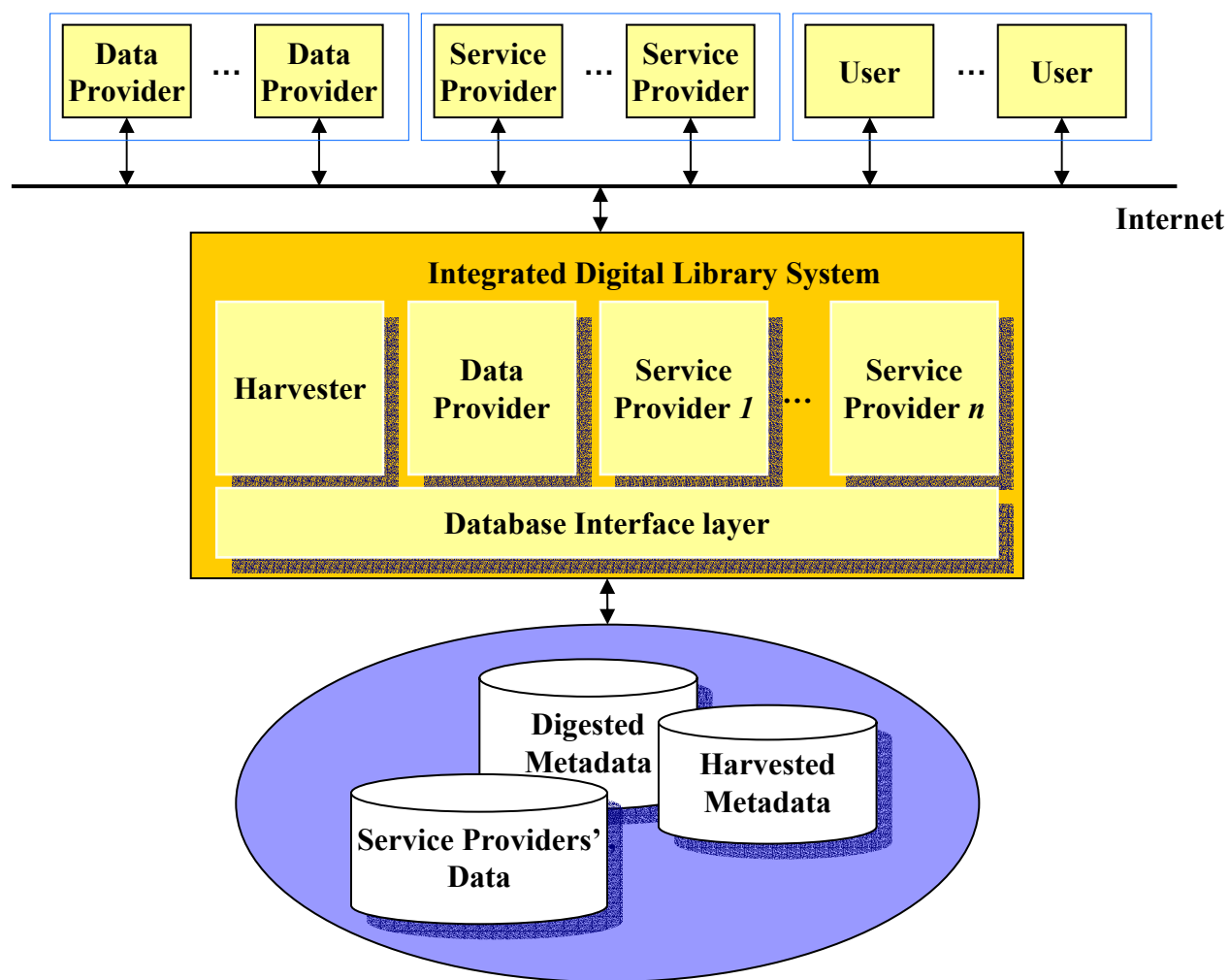
- Top-level clustering result

- 20x20 SOM
- 49 concept regions identified



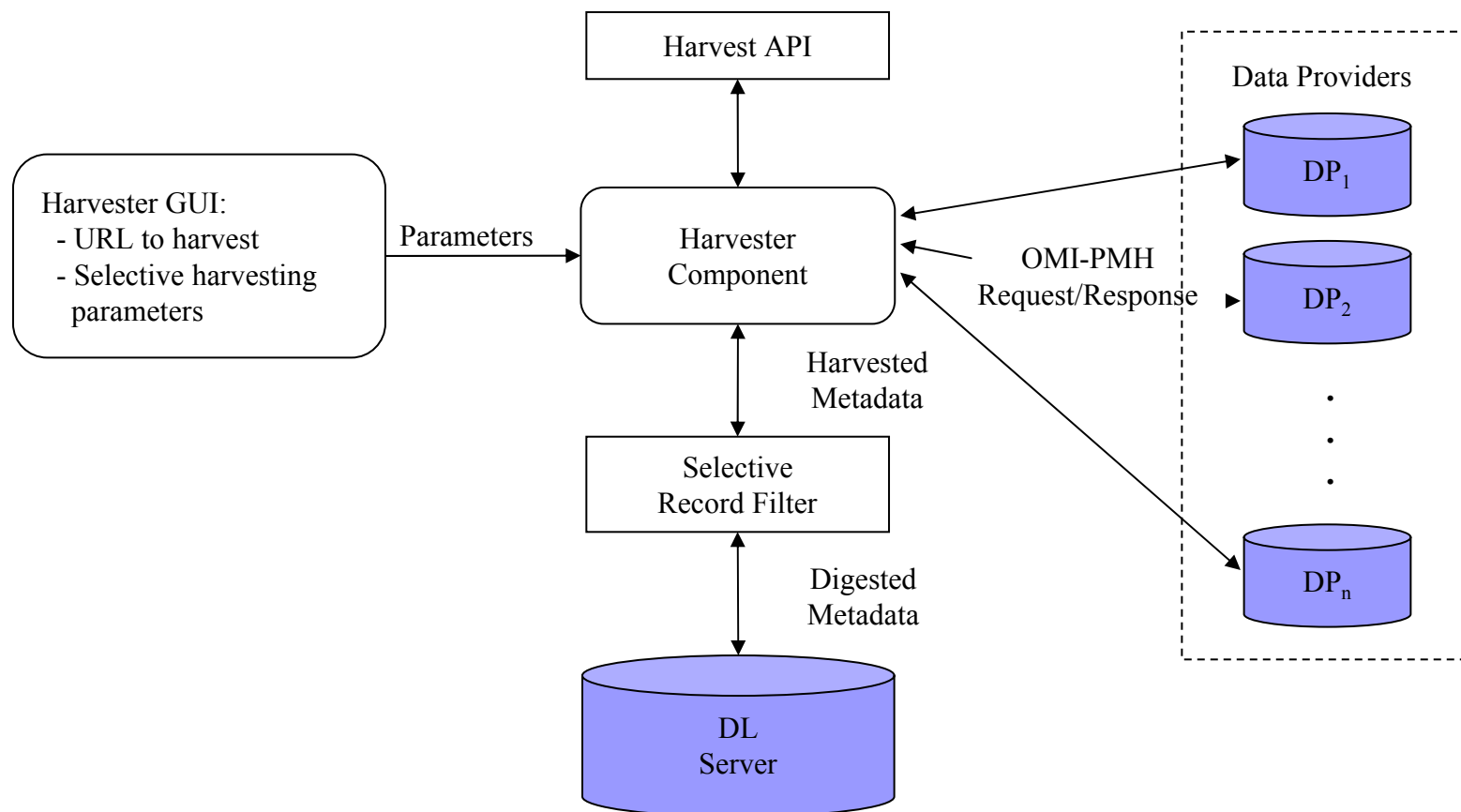
System Overview (1/5)

■ System Architecture



System Overview (2/5)

■ Harvester





System Overview (3/5)

■ Data Provider

- Expose single or combined metadata sets harvested to other harvesters
- Reformat metadata from different data providers to be harvested by other service providers (e.g., originally Dublin Core, reformat to MARC before exposing)

■ OAI-PMH Extension

- Allow service providers to harvest metadata sets from several data providers in a single request
- Example
 - OAI-PMH set parameter: “*set=LSUETD, VTETD*”

System Overview (4/5)

- Service Providers
 - Cross-archive search

Simple Search

Advanced Search

Filter Elements

University of Florida Digital Library Project

Concept: network

Back to Home

Displaying Records 1 - 10 of 100
Result Pages: 1 2 3 4 5 6 7 8 9 10 Next >>

- [Haptic Interaction with Three-Dimensional Bimapped Virtual Environments](#)
... CAM simulations. Methods for coping with **network** and computational latencies are described, and example applications are evaluated to explore the effectiveness of the system. This system can also provide an intuitive user interface for manipulating data within ...
Creator: Floyd, Jared Source: M.I.T. Theses
[More Info](#) [View Record](#)
- [Strictly Non-blocking WDM Cross-connects](#)
Using wavelength Division Multiplexing (WDM) technology, an optical **network** can route multiple signals simultaneously along a single optical fiber by encoding each signal on its own wavelength. If the **network** contains places where multiple fibers connect together ...
Creator: Rasala, April Source: M.I.T. Theses
[More Info](#) [View Record](#)
- [Physical Layer DSP Design of a Wireless Gigabit/s Indoor LAN](#)
The Wireless Gigabit/s Local-Area **Network** (WGLAN) project is aimed at providing high-speed data transmission between the Next Generation Internet and end-use devices within the home or office environment. The design of the digital signal processing (DSP) required at ...
Creator: Arriola, Esteban Clemente Source: M.I.T. Theses
[More Info](#) [View Record](#)
- [On the invariant impedance function and its associated group of networks](#)
No description provided

System Overview (5/5)

- Service Providers
 - Interactive concept browsing

Concept Harvester using the Self Organizing Map

The screenshot shows a web browser window titled "Concept Browsing - Mozilla (Build ID: 2002052918)". The address bar shows the URL "http://dinosaur.lite.cise.ufl.edu:8080/oaicb/index.html". The page header includes the University of Florida Digital Library Project logo and navigation links for FLORIDA, ILLINOIS, VIRGINIA TECH, and EDUCATION. The main content area displays a "Concept Harvester" window with a self-organizing map (SOM) of concepts. The map is a grid of colored rectangles representing different concepts, with labels such as "engineering", "computer engineering", "mechanical engineering", "architecture", "computer science", "landscape architecture", "environmental engineering", "sociology", "english", "geology", "mechanics", "art", "grammar", "forestry", "vocational education", "fisheries", "psychology", "petroleum engineering", "biotechnology", "diary", "food science", "chemistry", and "geography". A sidebar on the left lists various concepts with their associated scores, such as "aerospace (0.25)", "agribusiness (1.0)", "air traffic control (1.25)", "architecture (4.25)", "art (6.5)", "business administration (0.25)", "chemical engineering (2.25)", "chemistry (3.25)", "civil engineering (0.5)", "coastal sciences (1.75)", "collision (0.25)", "computer engineering (4.25)", "computer science (3.75)", "dairy (1.0)", "data transmission system (0.25)", "economics (2.0)", "education (0.25)", "engineering (1.25)", "english (6.0)", "environmental engineering (4.75)", "finite element method (0.25)", "fisheries (0.25)", "food science (0.25)", "forestry (1.0)", and "geography (5.25)".

Concept: network

The screenshot shows the same web browser window as the previous one, but with the "Concept: network" selected. The main content area displays a list of records related to the concept "network". The list includes the following items:

- [Haptic Interaction with Three-Dimensional Bitmapped Virtual Environments](#)
... CAM simulations. Methods for coping with **network** and computational latencies are described, and example applications are evaluated to explore the effectiveness of the system. This system can also provide an intuitive user interface for manipulating data within ...
Creator: Floyds, Jared Source: M.I.T. Theses
[\[More Info\]](#) [\[View Record\]](#)
- [Strictly Non-blocking WDM Cross-connects](#)
Using wavelength Division Multiplexing (WDM) technology, an optical **network** can route multiple signals simultaneously along a single optical fiber by encoding each signal on its own wavelength. If the **network** contains places where multiple fibers connect together ...
Creator: Rasilala, Ajmal Source: M.I.T. Theses
[\[More Info\]](#) [\[View Record\]](#)
- [Physical Layer DSP Design of a Wireless Gigabit's Indoor LAN](#)
The Wireless Gigabit's Local-Area **Network** (WGLAN) project is aimed at providing high-speed data transmission between the Next Generation Internet and end-use devices within the home or office environment. The design of the digital signal processing (DSP) required at ...
Creator: Arvelo, Eladio Clemente Source: M.I.T. Theses
[\[More Info\]](#) [\[View Record\]](#)
- [On the invariant impedance function and its associated group of networks](#)
No description provided

The sidebar on the left shows a list of concepts with their scores, including "civil engineering (0.5)", "coastal sciences (1.75)", "collision (0.25)", "computer engineering (4.25)", "computer science (3.75)", "algorithm (8.16)", "approximation (2.04)", "bias (2.04)", "compiler (10.2)", "design (2.04)", "environment (12.04)", "machine (6.12)", "mechanism (2.04)", "multicomputer (8.16)", "network (2.04)", "performance (4.08)", "programming (2.04)", "study (16.37)", "thesis (16.33)", "visi (4.08)", "dairy (1.0)", "data transmission system (0.25)", "economics (2.0)", "education (0.25)", "engineering (1.25)", and "english (6.0)".



Lessons Learned

- Metadata quality affects the usefulness and value of service providers.
- Many data providers could not be harvested due to various reasons: connection refused, service temporarily unavailable, and XML encoding errors.
- To be more efficient, our system may be improved in several directions.
 - The SOM processing for new or modified metadata is not feasible.
 - The size and lattice type of the SOM should be predetermined.
 - User evaluation is required.



Conclusion

- We have proposed the integrated DL system that integrates cross-archive search with data mining.
- We have also proposed the hierarchical SOM algorithm with two feature sets for clustering Dublin Core metadata.
- Our future researches are to evaluate the retrieval performance and to do user study.