A Distributed Digital Library of Mathematical Monographs: Technical Aspects of the CGM Protocol

David Ruddy
Cornell University Library

DLF—Spring Forum 2004

Project overview

- Interoperability: provide unified access to a distributed body of work
- Content: mathematics monographs (~2000)
  - University of Michigan Library
  - Cornell University Library
  - State and University Library Göttingen
- Three distinct, local systems
  - DLXS, DPubS, Agora
- Funding
  - NSF & DFG (Nov 2000 – Oct 2003)

Interoperability goals

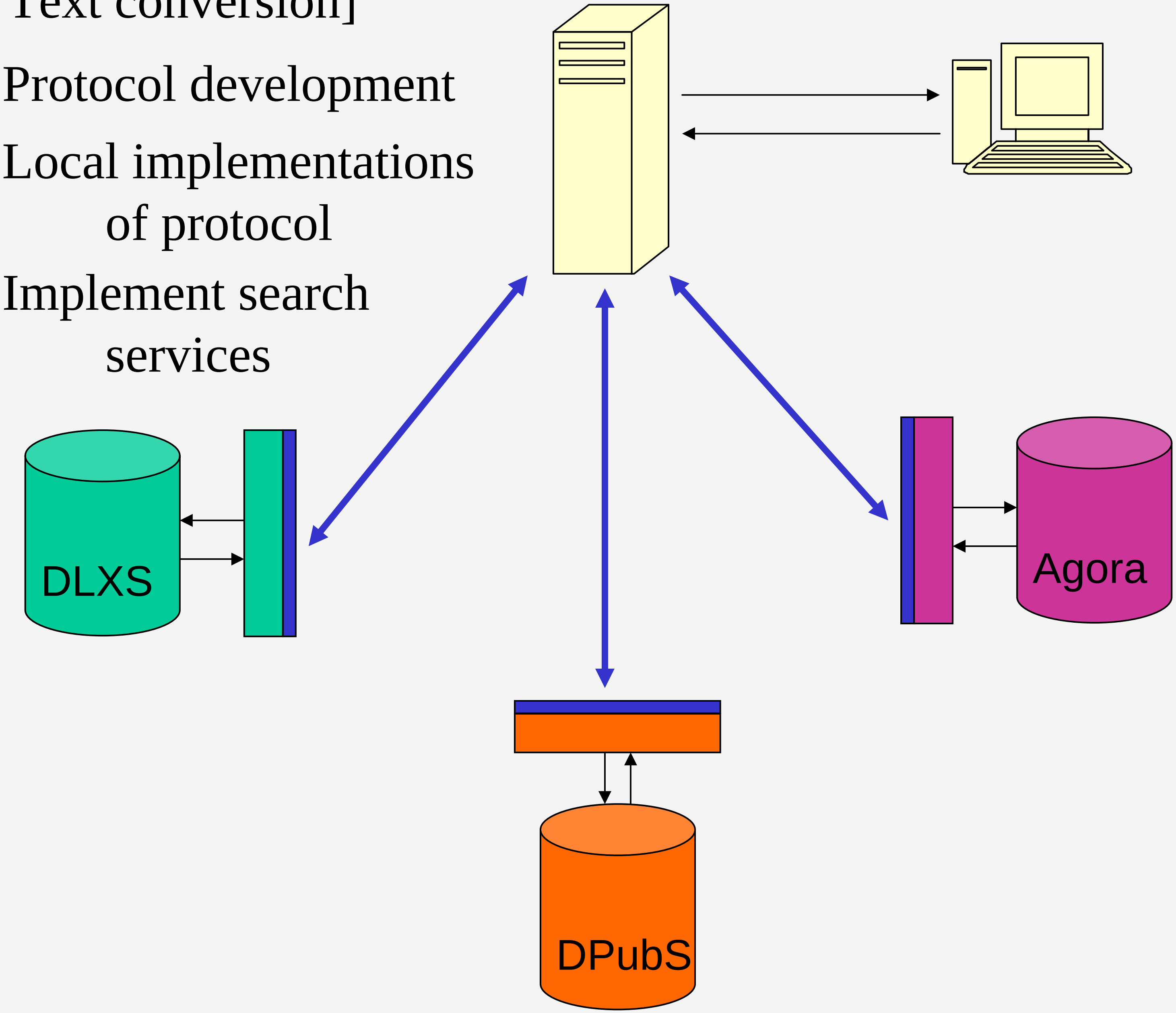
- Explore distributed search (vs. harvesting) as a federating technique
- Focus on full-text (vs. metadata)
- Communicate information about document structure among heterogeneous systems
- Create a protocol that abstracted search query and document structure information

[1. Text conversion]

2. Protocol development

3. Local implementations of protocol

4. Implement search services



Protocol goals

- Allow a user to search full-text documents in remote repositories
- Allow a service to retrieve information about a document and its structure, potentially as a navigational aid for users
- Allow a service to retrieve a full-text document or its component parts

Protocol ancestry

- Dienst (mid-90s)
  - Developed to support a distributed services digital library model
  - 27 verbs, 5 services
- Open Archives Initiative (Jan 2001)
- Our protocol (**CGM**) defined 10 verbs
  - 8 inherited from Dienst and revised
    - Removed all metadata communication
    - Extensive reworking of 3 verbs
  - 2 new verbs
  - Assumes OAI compliance

Ten CGM verbs

ListVerbs DescribeVerb	Retrieving information from a repository about its CGM implementation
Search	Document discovery
Terms	Retrieving document IP rights information
ListVersions ListViews Structure Formats	Retrieving information about document versions, structures, and distributable formats
Display Disseminate	Displaying or requesting documents or document components

Displaying or delivering docs

- Assume we have discovered a document
- We can:
  - Ask the local repository to display it
  - Ask the local repository for structural information about it, with the intent of presenting it in some meaningful way to users

Display verb request

```
http://some.cgm.server/script?protocol=CGM&
     verb=Display&ver=1.0&
     identifier=cul.math/00640001
```

General syntax (ListViews)

Request:

```
http://some.cgm.server/script?protocol=CGM
  &verb=ListViews&ver=1.0&identifier=cul.math/00640001
```

Response:

```
<?xml version="1.0" encoding="UTF-8"?>
<CGM>
  <responseDate>2002-10-02T19:20:30Z</responseDate>
  <request verb="ListViews" ver="1.0“
       identifier="cul.math/00640001">
       http://some.cgm.server/script</request>
  <ListViews ver="1.0">
    <identifier value="cul.math/00640001">
    <view id="v123-a" label="Page List" default="1" />
    <view id="v123-b" label="Chapters and Sections" />
  </ListViews>
</CGM>
```

Simple Structure verb response

```
<Structure ver="1.0">
  <identifier value="cul.math/00640001"/>
  <view id="v123-a" label="Page List" default="1">
    <div id="a123" type="maindocument" order="1“
               label="Entire Monograph" diss="0">
      <div id="a123-1" type="page" order="1" label="Page NA“ diss="1">
      <div id="a123-2" type="page" order="2" label="Page i“ diss="1">
      <div id="a123-3" type="page" order="3" label="Page ii diss="1"">
      <div id="a123-4" type="page" order="4" label="Page 1“ diss="1">
      <div id="a123-5" type="page" order="5" label="Page 2“ diss="1">
      <div id="a123-6" type="page" order="6" label="Page 3“ diss="1">
        . . .
      <div id="a123-251" type="page" order=“251" label="Page 248“
               diss="1">
    </div>
  </view>
</Structure>
```

More complex example

Simple Formats verb response

```
<Formats ver="1.0">
  <identifier value="cul.math/00640001"/>
  <divReq id="a123-4" type="page" label="Page 1">
    <format type=“gif" mime="application/gif“
            size=“24292" label="Page Image"/>
    <format type=“pdf.600" mime="application/pdf“
            size="55674" label="Page in PDF"/>
    <format type="ascii" mime="text/plain“
            size="3995" label="Un-proofed OCR text"/>
  </divReq>
</Formats>
```

Disseminate verb request

```
http://some.cgm.server/script?protocol=CGM&
     verb=Disseminate&ver=1.0&
     identifier=cul.math/00640001&
     div=a123-4&
     format-type=gif
```

A search using CGM

- A user formulates a search within a search service
- Search service translates that query into a CGM request; broadcasts that request to relevant repositories
- Repositories translate the request into local query syntax and perform the search
- Repositories package search results into CGM compliant responses and return to search service
- Search service translates results for user

Search verb request

```
http://some.cgm.server/script?protocol=CGM&
     verb=Search&ver=1.0&
     set=math&
     resultSize=100&
     field1=author&value1=todhunter&
     field2=fulltext&value2=trigonometry&
     op2=and
```

Additional arguments available:

```
     sort
     startResult
```

Working search services

- Two search services in production:
  - Michigan
    http://www.hti.umich.edu/m/mathall
  - Cornell
    http://mathbooks.library.cornell.edu

Possible search enhancements

- Improving precision
  - Problem of dissimilar structures
  - Abstracting document structure
- Search service improvements
  - More sophisticated handling of search results to improve performance (scaffolding techniques)
- Support metadata types in search results

Improving precision

- Take more advantage of document structure to create more sophisticated queries
  - Word A and B on the same page, in the same chapter, paragraph, etc.
- Yet mediating dissimilar structures in query construction is expensive, and likely unsatisfying to the user
  - Query possibilities will change depending on the mix of repositories

Abstracting document structure

- Define abstract document structure
- Repositories map local document structures to this abstract model

Possible abstract doc model

maindocument [required]	the entire document
div-high	structures such as TEI front, body, back
div-mid	chapters, sections, miscellaneous divisions
div-low [required]	page, paragraph, illustrations, charts, etc.

Some advantages

- Works around the problem of dissimilar document structures
- Makes explicit assumptions about child/parent relationships that can then be used in query construction
- Allows those closest to the documents to translate specific document structures to an abstract model

Search service improvements

- Enhance query performance with more sophisticated results negotiation, delivery, and display
  - More actively and creatively mediating the search transaction

Support metadata types in results

- Search result records are now returned in a fixed metadata format
- Allowing metadataPrefix selection with Search request to support:
  - Richer search result display possibilities
  - Selective harvesting
- Potential cost: expensive transaction

CGM strengths

- Specialized functional scope
  - Searching distributed repositories of full-text
  - Conveying info about document structure
  - Requesting documents or document components
- Designed to work alongside other specialized protocols (OAI)

  ```
  protocol=CGM
  ```
- Flexible development path
  - Verb versioning

More information

- Project web site
  - http://www.library.cornell.edu/mathbooks
- Search Services:
  - Michigan

    http://www.hti.umich.edu/m/mathall
  - Cornell

    http://mathbooks.library.cornell.edu