The Digital Library at Penn

Developing Integrated Digital Collections in a Distributed Setting

Delphine Khanna
Digital Projects Librarian
University of Pennsylvania Library

DLF Forum on Digital Library Practices
April 1, 2000

Digital library activities at Penn: Initial development model

- Distributed across a number of units in the Library.
- Developed in autonomous manner by each unit.
- Developed in *ad hoc* fashion:

  Sometimes use idiosyncratic methods.

Digital library activities at Penn: Initial development model (2)

- Some of the specific Library units involved:

  SCETI, Special Collections, Fine Arts Slide Collection, South Asian Collection, Music Collection, Systems, Public Services, Collection Development.

- Projects vary from small to large.

  Fine Arts Slide collection, SCETI, OUP/Penn History Books Project.

- Some are actual projects, some are part of routine operations (e-reserves, acquisition of e-journals…).

Limits of the model

- Very valuable activities,
- But a number of problems have emerged:
  - Lack of consistency across collections,
  - Difficulty to scale up, to maintain some of the collections,
  - Accessibility is limited.

    (E.g., not all collections are satisfactorily searchable.)

New model

- Goal: to develop the Digital Library in a more integrated manner.
- Relatively recent.
- Summer ’99, 3 Digital Library specialists hired to support this process:
  - John Ockerbloom, Digital Library Architect,
    Ph.D. in Computer Science,
  - Delphine Khanna, Digital Projects Librarian,
    Master’s in Linguistics and Computer Science and M.L.S.,
  - Mike Winkler, Web Manager,
    M.L.S., IT background,
    Pre-existing position, but with different orientation.

New Model (2)

- 2 persons in Systems, 1 in Public Services.
- Work very closely with each others.
- Supplement people already in place in various Library units.

Integration of Digital Library Collections

- Targets all types of resources indistinctly:
  - Acquired and licensed,
  - Produced in-house,
  - Includes integration of e-resources with print collections.

Balance

- Strike the balance between:
  - Distributed operations, and
  - Integration effort.
- Distributed activities:
  - Foster creativity, flexibility, specialized technical skills.
- Coordination and integration:
  - Crucial for development of large-scale Digital Library.

How do we work?

We share our time between:

- Working with specific Library units,
- Working on improving the overall architecture.

Working with specific Library units

- Offer consulting:
  - discuss specific projects,
  - suggest possible tools and techniques.
- Develop applications:
  - Prototypes,
  - Selected production applications.
- Offer training, information sessions as needs arise.
- For instance: Fine Arts Slide Collection

Working on the overall architecture

Participation in administrative bodies

- Strategic Planning Group:
  - Key senior Library administrators and the 3 Digital Library specialists,
  - Provide guidance and direction for the overall effort of Digital Library development,
  - Sets overall priorities.
- Web Advisory Group:
  - Concerned specifically with the Library Web site.
- Department Heads Meetings (informative role).

Research,
Exploration of new technologies

For instance:

- Digital preservation issues,
- Citation resolution,
- Data provenance.

Developing generic tools and techniques

- Core of our work:
  - What will really make integration happen.
- Integration can’t just happen in a top-down manner, by decree:
  - Projects are ongoing operations,
  - We can’t just arrive and change everything.

Strategy

- Identify various tools and techniques,
  - Which can be used in all Digital Library projects at Penn,
  - Which have the potential to improve the overall quality and integration of the Digital Library collection.
- Make those tools available.
- Publicize them to various Library units, who will choose to adopt them at their own pace.
- Goal:
  - Build sounder foundations for our Digital Library to make it more scalable.
  - All this without disturbing too much the current operations.

A few examples of techniques/tools that we are implementing

Universal Resource Identifiers

- Currently, we are implementing a handle server.
- Very important for the implementation of everything else.
  - To experiment with delivery technologies:

    Need to be able to modify URLs of resources.
  - For instance:

    From static HTML to database-driven dynamic HTML.
  - Currently links break if any change occurs:

    Most importantly links in MARC records break (856 fields).
- For resources developed in house as well as acquired/licensed.

MrSID

- Currently, inconsistency in quality among e-image collections.
- Towards unified model:
  - TIFF Master files,
  - More flexibility at delivery time.
- Currently implementing MrSID as a delivery technology:
  - Proprietary technology,
  - High compression rate, deliver JPEGs on the fly,
  - Zooming capabilities,
  - Users can specify the size of the JPEG images.

Implementing MrSID: two advantages

- Incentive to produce high resolution TIFF images.
- More flexible delivery.
  - Renaissance in Context Project:
    - Access by scholars vs. high-school students.
  - Fine Arts Slide Collection:
    - Access by students from home,
    - Projection on wall screens in the classrooms.
- Note: TIFFs are kept as master files:
    - So proprietary format is not a problem.

MrSID:
Broad range of applications

Suited for many types of images:

- Art slides,
- Manuscript facsimiles,
- Maps,
- GIS,
- etc.

Repurposing of MARC records

- Another set of techniques that we are exploring currently.
- General idea: we have been under-using our MARC records,

  because we let OPAC vendors define what we should do with them.
- MARC in itself relatively flexible,

  e.g. support of hierarchical relationships (773 field).

Beyond the OPAC interface

- OPAC architecture at Penn:
  - Endeavor's Voyager,
  - Runs on Oracle.
- We can do much more with MARC:
  - If we develop alternate Web interfaces connecting directly to the underlying database.

Example

- E-journal pages on Library Web site:
  - Over 2300 e-journals listed,
  - Previously hand-coded in HTML:

    Process completely separated from the cataloging of e-journals in the OPAC.
  - Now: use e-journals' MARC records to generate list of e-journals on the fly.
- Many other potential applications.

Why is it appealing to do more with our MARC records?

- Avoids duplication of efforts.

  For resources which will be cataloged in the OPAC anyway.

- The OPAC is our traditional 1-point access tool.

  => A good starting point to develop integration of all resources (digital and non-digital).

Why is it appealing to do more with our MARC records? (2)

From an organizational standpoint:

- MARC already used across the entire library, well accepted,
- A lot of in-house expertise,
- Workflow already organized around MARC,
- => Relatively easy to implement new uses of MARC.

What we have learned
(or *are learning*)

Immediate Benefits

- By definition, specific projects are very focused on their specific goals.
- In practice, abstract notion of integration: not a priority.
- => Good strategy: offer generic tools that
  - Will improve the overall integration and consistency of the Digital Library,

    AND
  - Have an **immediate added value** for projects.

A tool which is ready and available is worth a 1000 words

- *A priori*, Library units ask for custom-designed solutions:
  - Highly focused on the unique aspects of their projects,
  - So talking about generic solutions in theory is ineffective.
- But once the generic tools are in place, adoption is much easier.

  => Use prototypes as a basis for discussion.

Modularity is great

- “Tool” approach naturally modular.
- Helps to not overwhelm people with change:
  - Easier to “put your arms around” each tool,
  - People can set their own implementation schedule,
  - They keep a better control over the development of their projects.
- As opposed to:
  - “You should change the entire architecture of your project.”

A lot of informal, semi-formal discussions

- Dynamic, interactive process:
  - Talking to people to understand specific needs,
  - Identifying common needs across projects,
  - Exploring new solutions.

- Build common grounds,
- Enable knowledge transfer,
- Develop trust.

Be on the lookout for cross-project solutions

- Work on each project with other projects in mind.
- Use each project as a pilot:
  - Show it to other Library units,
  - It will give them new ideas for their own projects.

Challenge

Avoid getting sucked into specific projects.

- Make sure to keep enough time to work on general issues.

Conclusion

Supporting the overall integration of the Digital Library:

- A number of challenges,
- Model chosen by Penn is interesting,
- Very enriching experience.