Introduction

– Todd Carpenter, Managing Director, NISO

Data quality, policy, and large-scale data flows

How do regional consortia establish and implement policies to allow them to cope with increasing amounts of data in a widening variety of formats?

– Hilary Newman, Vice President of Implementation Services, Innovative Interfaces, Inc.

Academic library perspective

What are the successes, stress points, and failures from the perspective of the academic library?

– Maribeth Manoff, Coordinator for Networked Service Integration, University of Tennessee, Knoxville
– Robert McDonald, Associate Dean for Library Technologies, Indiana University

Agenda

• To Merge or Not To Merge?
• Challenges
• Quality Implementation Practices
• Strategic Policy Decisions
• Data Quality and Standards
• Project Management

My Background

• Work with libraries to merge and populate bibliographic and other system data
• 15 years at Innovative
• MLIS University of North Texas

To Merge or Not To Merge: Merge!

• Initial Data Population Planning
• Ongoing Maintenance and Growth
  – Authority Control
  – Outsourcing
  – Institutionalizing Practices
• Duplicate Records, Who Wins?
### Challenges
- Quality of Data or Cataloging Practices
- Maintaining Encoding Levels?
- Record Source Quality
- Local Practices and Needs
  - Item identification, barcodes, call numbers
  - Licensing Information
  - ISSN/ISBN Cataloging Practices
  - 856 and other URL Link Uses
  - Etc.…

### Quality Implementation Practices
- Following Standards
- Set Local Control Policies
- Match Points and Control Numbers
- Consider End User, Is More Data Better?
- User Interface Considerations

### Strategic Policy Decisions
- Consider Data Alive and Evolving
- Consistency and Simplicity Win
- Don’t Make Policies Based on Today’s Needs
  - User Interface Changes
  - Future Data Mergers
  - Possible Data Splits?
  - Staff Costs to Maintain Customized Practices

### Data Quality and Standards
- Cataloging Quality
- Bibliographic Records
- Electronic Resources
- Metadata
- Beware the Cleanup Project
- Leverage Computing Power

### Project Management
- Who Decides and Who Does the Work?
- Reconsider Local Workflows
- How Do Others Do It Successfully?
- Plan Ahead
- Don’t Plan for Perfection

### Conclusions
- Plan for the Life of the Data
- Standards!
- Decision Making and Teamwork
- Make Computing Power Work For You
- Be Realistic
Data Migration and System Population at the UT Libraries

- Large, “one-time” system migration
  - Implementation of a new Integrated Library System (ILS)
  - Ongoing migration / constant system population
    - The link resolver “knowledge base”
    - A “next generation” discovery and delivery system

Large, “One-Time” System Migration

- Migration to a new ILS (2003)
- Moved to Aleph from 3 separate systems:
  1) Cataloging and Circulation - Horizon
  2) Acquisitions and Serials - GEAC
  3) Web OPAC – OCLC SiteSearch (Z39.50)
- Data migrated was in known, or at least familiar, formats

Migrating Data in Known Formats

1. Bibliographic data in MARC format
   - 1.6 million bib records “straight” MARC to MARC
   - Brief order bib records from acquisitions system
     - Some matching and mapping required for orders to be attached to the correct bibliographic record
   - Course reserves records
     - Also in MARC format; stored in a different database table in the new system
2. Item / Holdings data
   - Translation algorithm for item records considering location, collection, and call number (from either bib or item) to holdings records in MARC format
   - Import data stream of item records from Horizon to create holdings and item records in Aleph

Known Formats (cont.)

- For Non-MARC data – formats are not standardized, but data elements within the tables are familiar
  - Circulation data from Horizon
    - Patrons (4 tables)
    - Loans
    - Hold requests
  - Acquisitions and Serials data from GEAC
    - Orders
    - Subscriptions
    - Budgets
    - Vendors, Vendor addresses
Procedure for ILS Population – Non-MARC Data

1. Export data from Horizon or GEAC into tab-delimited file
2. Set up local MySQL database with fields equivalent to database tables in Aleph
3. Import tab-delimited files into database, moving data elements into correct fields for Aleph
4. Import from local database into Aleph database tables

ILS Migration - The UT Experience

- Labor intensive and time consuming; required extreme and exclusive focus
- Export -> import -> export -> import process used with non-MARC data took a number of iterations to get right
- Familiar formats made it relatively easy to see when the pieces fell into the right places, a satisfying experience
- Internal workflows and some patron empowerment features (e.g. online renewal) were improved with one integrated system

ILS Migration - The UT Experience (cont.)

- One-time nature of the migration and familiarity with the data may have led to less innovation
- Indexing and display of MARC fields was mostly taken from previous system
- “We got it right” feeling persisted
- Didn’t have enough time to spend on the user experience side (the OPAC)
- Following completion of migration, changes that required reindexing were often postponed

Toward Ongoing Migration and New Data Formats

- UT has developed a number of ongoing processes and services moving data in and out of the Aleph system
  - Load patron information with exports from campus LDAP system
  - Export of invoices for ingest and processing in SAP
  - SQL queries for custom reporting
- Newer, more modular systems require ongoing migration of data in less familiar formats

The Link Resolver “Knowledge Base”

- Emphasis of OpenURL link resolver software (SFX from Ex Libris at UT) is on the user experience - simplifying access
- Ongoing data updates are essential for linking accuracy
- Software providers collect and distribute large amounts of data from electronic resource vendors – an invaluable service
- Processes for updating the KnowledgeBase to reflect current subscription information are equally important

Link Resolver KnowledgeBase (cont.)

- Data elements in the KnowledgeBase are keyed to access
- Constant vigilance in keeping the KnowledgeBase updated, and different data formats, helps us think in new ways, e.g.:
  - What are the most important data fields to enable efficient and effective searching for journals?
  - What is the best way to express relationships between serial titles?
The “Next Generation” Discovery and Delivery System

- Another module with a focus on the user experience (Primo from Ex Libris at UT)
- Catalog data is “transformed”
- MARC Export from ILS
- Import / transformation into an XML format
- Other data sources, e.g., records from an OAI repository, can be added

“Next Generation” Discovery and Delivery (cont.)

- XML data elements are keyed to discovery and delivery, e.g.,
- Facet fields for navigation of search results created from author, subject heading, classification, availability information, …
- Dedup and FRBR fields to facilitate streamlined display of duplicate or “like” records
- “Rules” for data transformation are understandable and configurable
- Reindexing does not require system downtime

“Next Generation” Discovery and Delivery (cont.)

- Ongoing data migration, into new and configurable data format, encourages innovation, e.g.: ∗ How much does “format” matter, and how does it play into faceted navigation? ∗ New ways to bring authority data into the mix without having to browse an index

The Challenge of Ongoing Data Migration

- A challenge in the best sense of the word, fostering new thinking and innovation
- Requires procedures that are both rigorous and flexible
- More to look forward to on this front, as other, modular, systems come online at the UT Libraries (e.g., ERMS)

Thank you!
Hello – Greetings - Welcome – Bonjour - Tag

Tagging
- #niso
- #data
- @mcdonald
- GoogleWave
  - with:public niso

Robert H. McDonald

OVERVIEW
- Institutional Background
- Data at IU
  - Discovery
  - Curation
  - Storage
  - Scale
- Data Futures
- Above Campus

Indiana University Libraries
- Indiana University
- Statewide Location, Indiana
- Large, Research Oriented Public University
  - 101,000+ students
- SIRSIDynix Symphony 3.2.1 Library Management System
  - 6.5 million items
- 9 Regional Campuses
- ILS Managed by Libraries/Enterprise Applications

DATA @IU
- Data Discovery
  - Bibliographic Data
- Data Curation
  - Local
  - Consortial
- Data Storage
  - Local
  - Consortial

DISCOVERY DATA
- SIRSIDynix Symphony Web OPAC
- WorldCat Local Discovery Interface
- Testing Ebsco Discovery Interface
- Mobile Discovery
  - Boopsie Mobile Search
  - MIT Mobile Framework (campus)
- Multiple Instances
  - Dspace – IU ScholarWorks
  - Open Journal System – IU ScholarWorks
  - Fedora (modern DLP infrastructure)
  - Legacy Digital Library Systems

LEGACY ILS
- Public User Interface:
- Legacy ILS Functionality
- Vendor Based Metadata Enhancement
- Data Stores:
- Staff Interfaces:
  - Circulation
  - Acquisitions
  - Cataloging
  - Serials
  - Functional modules:
Legacy ILS+Current Management Tools

- Learning Mgmt./Campus Portal Feeds
- Federated Search/Advanced Discovery
- OpenURL Linking
- Electronic Resource Mgmt. System
- Proxy/VPN for eContent Access
- Legacy ILS Functionality + eContent Module + Advanced Discovery Interface
- Unified Workflow Interface for Staff

Good Search – Decouples Discovery from ILS

- NextGen Search System/Faceted Search
  - exports and reformats MARC from ILS
  - Raw MARC data
  - Flat text files
  - Parse text files
  - Indices
  - Web 2.0 Engine
  - HTTP
  - User Interface

IU DISCOVERY MOBILE

- MIT MOBILE FRAMEWORK
- DATA APIs
- SAKAI
- IU KNOWLEDGEBASE
- IUCAT SEARCH

IUCAT DISCOVERY MOBILE

- BOOPSIE
- MARC RECORD EXTRACT
- AUTOMATED UPDATES
- IU INTELLIGENT INFRASTRUCTURE

SCHOLARWORKS DISCOVERY MOBILE

- BOOPSIE
- DSPACE
- OAI HARVESTING
- AUTOMATED UPDATES
- IU INTELLIGENT INFRASTRUCTURE
DATA DELUGE

- As volume of digital scientific data increases so increases opportunities for data-driven science
- Funding agencies beginning to mandate that research data products be made publically available
  - "The National Science Foundation is committed to the principle that the various forms of data collected with public funds belong in the public domain."

DATA CURATION

- IU Scholarly Data Services
  - IU ScholarWorks Data - Publishing Services
  - Consortial Data Curation/Publishing (HathiTrust)
  - Consortial Data Curation (CIC Research File System)

IU SCHOLARWORKS DATA

- IU ScholarWorks Data
- HTTP Server
- hpsfs filesystem
- IU MDSS (HPSS)
- MDSS web server

CONSORTIAL DATA - HATHITRUST

- Indianapolis
- Ann Arbor
- Isilon OneFS Currently Supports up to 2.3 PB between Two Nodes
- Two Node Federated Repository

CONSORTIAL DATA – CIC RESEARCH FILESYSTEM

- 7 Node Wide Area File System
  - Indiana University
  - Michigan State University
  - Ohio State University
  - Penn State University
  - University of Illinois
  - University of Iowa
  - University of Minnesota
- Fedora Curation Services
- GPFS-WAN – Global Parallel File System Wide Area Network
- SoFS – Scale Out File System
Above-Campus Services: Shaping the Promise of Cloud Computing for Higher Education
Brad Wheeler and Shelton Waggener – Educause Review 44(6)

DISCUSSION

Questions?

All questions will be posted with presenter answers on the NISO website following the webinar: