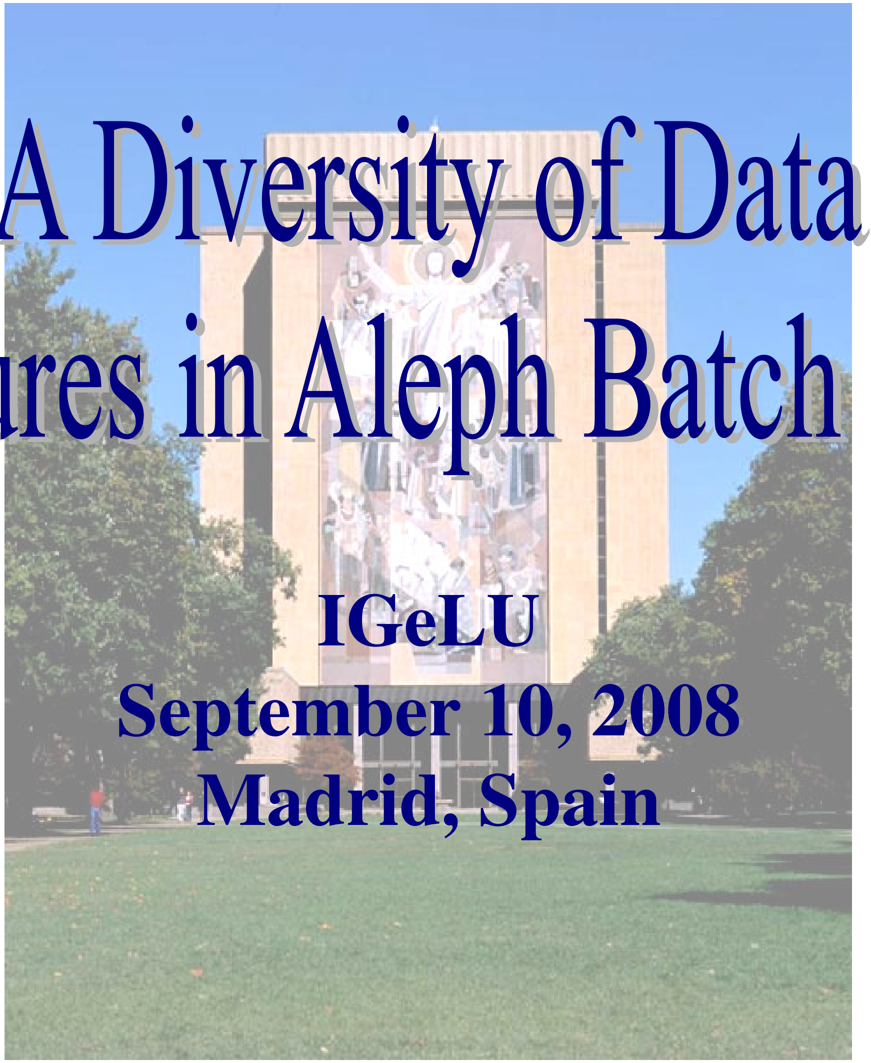


A Diversity of Data: Adventures in Aleph Batch Loading

IGeLU
September 10, 2008
Madrid, Spain

UNIVERSITY OF NOTRE DAME

Introduction

- Pascal Calarco, Head, Library Systems Department, pcalarco@nd.edu

- Mark Dehmlow, Electronic Services Librarian, mdehmlow@nd.edu

MARC Batch Loading Working Group

- Phil Andrzejewski, Library Systems
- Pascal Calarco, Library Systems
- Mark Dehmlow, Electronic Resources
- Mary Lehman, Electronic Resources
- Eric Morgan, Digital Access & Information Architecture

Overview

- Goal: Create some discussion about batch loading now and what it could be
- Context and background
- Batch Loading and Limitations of tab_fix et. al.
- MARC
- Programming the way forward?
- The Future

“Its About the *Information*, Marty”

- *Sneakers (1992)*

Data Slinging: Core Business

- Activity shifted from Aleph functionality implementation/tweaking to all sorts of data automation
- Reporting: data marts, metrics, decision support
- MARCIVE: gov docs, authority processing
- Vendor records for microform sets, electronic products
- EDI: order, pay; multiple vendors, differences
- We spend a lot of time moving records around
  - How efficient is this?
  - Is there a better way?

Drivers: Title-level access to digital collections

- Large digital collections (EEBO, ECCO, etc.)
- eBooks
- E-Journals: SFX MARCIt services
- Newspaper collections
- Deep Historical collections:
  - Poetry, literature and drama collections
  - Area studies & subject collections
  - Map collections
  - Parliamentary papers & government collections

Discovery vs Inventory

- Most patron-based discovery/access only requires a handful of data points:
  - title, uniform title
  - subject
  - call number
  - URL

- Some extra data useful for bibliographies
  - publication information for books,

- Most data in a MARC record is administrative – it helps us distinguish items from each other, and keep tabs on publication frequencies, basic content and physical descriptions, where journals are indexed, record identifies, etc.
  - North American Title Count

Collections Access: Silos

- Challenges:
  - Proliferation of silos of information
  - Barriers to discovery and access
  - Opening rich digital collections to undergraduates, non-specialists
  - Variability of MARC quality and availability

Aleph Functionality Configuration

- tab_match: define match points for manage_36 (eg. 035, 001, etc.)
- tab_merge: define merging routine to be used
- tab_fix: apply optional changes to incoming records; semi-scriptable
- tab_match_merge: match & merge routines & weights
- tab_weights: weighting parameters for matching
- tab_merge_overlay: defines fields to be overlaid/retained during loading
- tab_z30, tab_mapping: create holdings records for MARCIVE loading

Aleph MARC Loading: Steps

- Hasn't changed substantially since 2003
- Check input file against database (p_manage_36)
- Load catalog records (p_manage_18)
- Modify MARC records (p_file_08)
- Create Holdings and Item Records (p_manage_50)

Limitations of Fix Routines/Global Changes

- Add/delete an entire field
- Copy one field to another
- Cannot handle manipulation of 008 data
- Cannot handle complex routines at subfield level
- Global changes have little control over field load location (adds)
- How well does this scale?

Batch Loading

- Loading itself has few issues
  - # of records, batches 10,000+ starts to slow down indexing ~ workflow for staff
- Most issues in preparation and management
  - Need records to fit within local cataloging practices (change over time)
    - Local bases, format
  - Need way to recognize, search for, pull records out for addition, updating, deleting
- Consortial issues and ADM records

Challenges to Batch Loading Large Sets

- Catalog is “gentile” – requires many fields to have very specific formatting

- Variability in quality for vendor provided records
  - ➢ Records created by vendors
    - Black Drama authorities
  - ➢ Records repurposed from print

- Difficulty multiplies with holdings (n extra records)

- Staffing considerations
  - ➢ Some institutions have a dedicated person just for this purpose

The MARC Record

```
01259cam  2200301 a
4500005001700000006001900017007001500036008004100051020003300092020001600125035002300141040002100
1641000021001852450079002062600055002853000048003404400052003885040053004405060051004935330097005
445990018006415990019006596500032006786500044007107100024007547300036007788560143008 14□2007081415
3553.0□m^^^^^^^^d^^^^^^^^□cr un ^^^uuuua□030516s2003^^^^nyuabo^^sb^^^^001^0^eng^^□
□a0684314290 (electronic book)□  □z0684312573 □  □a(MiFhGG)0684314290□  □aDLC□cDLC□dMiFhGG□1
□aStepto, Gabriel.□14□aThe African American years□h[electronic resource] /□cGabriel Burns Stepto.□
□aNew York, N.Y. :□bCharles Scribner's Sons,□cc2003.□  □aix, 475 p. :□bill., maps, photos. ;□c29 cm.□
0□aChronologies of American history and experience□  □aIncludes bibliographical references and indexes.□
□aAccess restricted to subscribing institutions.□  □aElectronic reproduction.□bFarmington Hills,
Mi:□cGale,□d2003.□nAvailable via World Wide Web.□99□anew□d20080314□99□aauth□d20080314□
0□aAfrican Americans□xHistory.□ 0□aAfrican Americans□xHistory□vChronology.□2 □aThomson Gale
(Firm)□0 □aGale virtual reference library.□40□uhttp://lib-
[illegible]r?url=http://galenet.galegroup.com/servlet/eBooks?ste=2[illegible]&docNum=CX3[illegible]0199999&c=[illegible]d[illegible]ref
[illegible]F ACCESS□□
```

PERL Automation or The “Brute Force” Approach

- The origins for this work came out of our MARCIt project

- Two processes:
  - Analysis script outputs field frequencies, information about what kind of record it is
  - Scripts are then written to make changes resulting from analysis

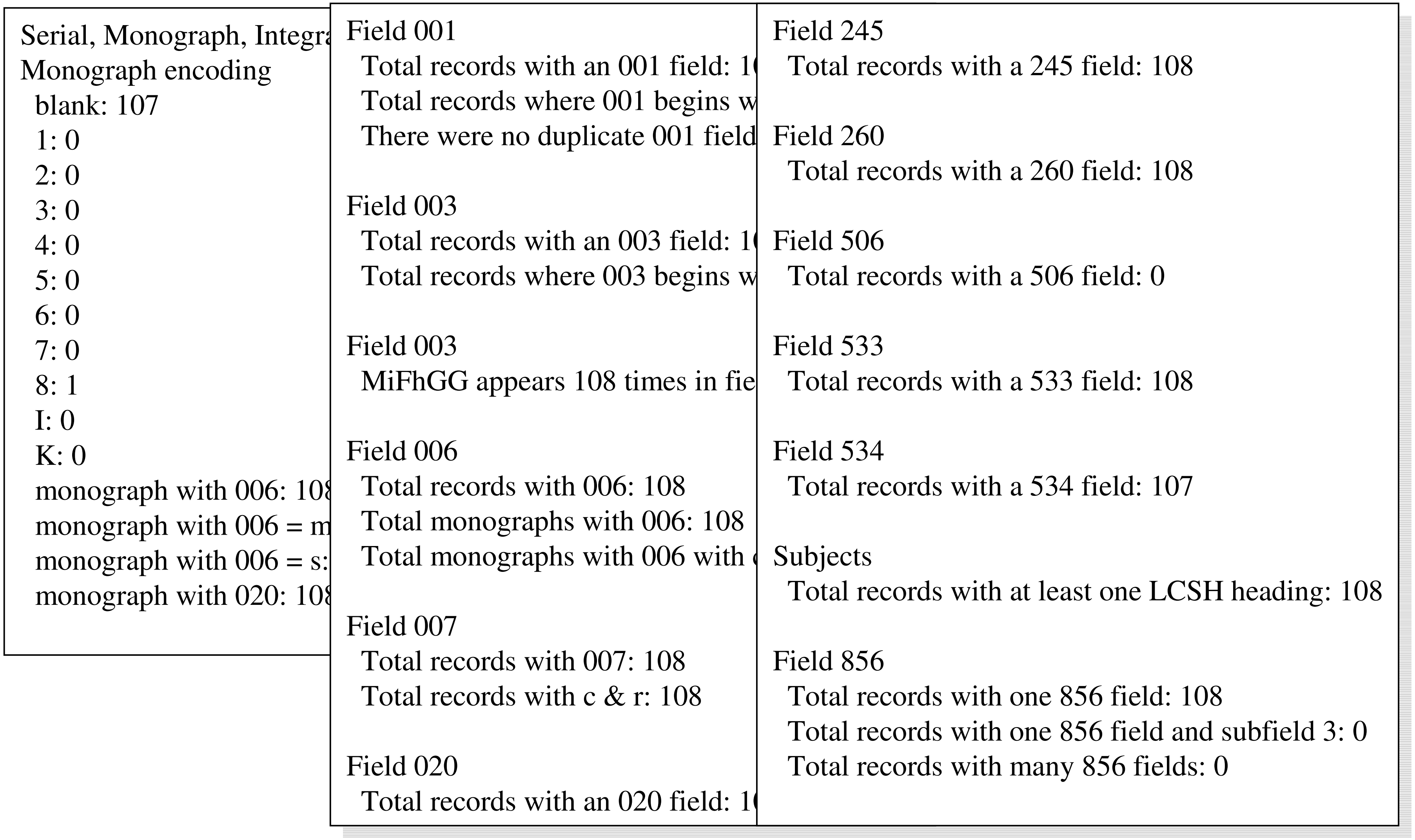
Programmatic MARC Record Manipulation

```
01259cam  2200301 a
4500005001700000006001900017007001500036008004100051020003300092020001600125035002300141040002100164100002100185245007900206260005500285300004800340440005200388504005300440506005100493533009700544599001800641599001900659650003200678650004400710710002400754730003600778856014300814□20070814153553.0□m^^^^^^^^d^^^^^^^^□cr
un ^^^uuuua□030516s2003^^^^nyuabo^^sb^^^^001^0^eng^^□  □a0684314290 (electronic
book)□  □z0684312573 □  □a(MiFhGG)0684314290□  □aDLC□cDLC□dMiFhGG□1 □aStepto,
Gabriel.□14□aThe African American years□h[electronic resource] /□cGabriel Burns Stepto.□
□aNew York, N.Y. :□bCharles Scribner's Sons,□cc2003.□  □aix, 475 p. :□bill., maps, photos.
;□c29 cm.□ 0□aChronologies of American history and experience□  □aIncludes bibliographical
references and indexes.□  □aAccess restricted to subscribing institutions.□  □aElectronic
reproduction.□bFarmington Hills, Mi:□cGale,□d2003.□nAvailable via World Wide
Web.□99□anew□d20080314□99□aauth□d20080314□ 0□aAfrican Americans□xHistory.□
0□aAfrican Americans□xHistory□vChronology.□2 □aThomson Gale (Firm)□0 □aGale virtual
reference library.□40□uhttp://lib-
proxy.nd.edu/login?url=http://galenet.galegroup.com/servlet/eBooks?ste=22&docNum=CX340919
9999&q=nd_ref□yClick for ONLINE ACCESS□□
```

```
LDR 01277cam  2200301 a 4500
005    20070814153553.0
006    m       d
007    cr un    uuuua
008    030516s2003    nyuabo  sb    001 0 eng
020    _a0684314290 (electronic book)
       _z0684312573
035    _a(MiFhGG)0684314290
040    _aDLC
       _cDLC
       _dMiFhGG
050 00 _aE185
       _b.S797 2003
100 1  _aStepto, Gabriel.
245 14 _aThe African American years
       _h[electronic resource] /
       _cGabriel Burns Stepto.
260    _aNew York, N.Y. :
       _bCharles Scribner's Sons,
       _cc2003.
300    _aix, 475 p. :
       _bill., maps, photos. ;
       _c29 cm.
440 0  _aChronologies of American history and experience
504    _aIncludes bibliographical references and indexes.
506    _aAccess restricted to subscribing institutions.
533    _aElectronic reproduction.
       _bFarmington Hills, Mi:
       _cGale,
       _d2003.
       _nAvailable via World Wide Web.
599 99 _anew
       _d20080123
599 99 _aauth
       _d20080123
650  0 _aAfrican Americans
       _xHistory.
650  0 _aAfrican Americans
       _xHistory
       _vChronology.
710 2  _aThomson Gale (Firm)
```

```
01277cam  2200301 a
45000050017000000060019000170070015000360080041000510200045000920350023001370400
02100160050002200181100002100203245007900224260005500303300004800358440005200406
50400530045850600510051153300970056259900180065959900190067765000320069665000440
0728710002400772730003600796856014300832□20070814153553.0□m        d        □cr un
uuuua□030516s2003    nyuabo  sb    001 0 eng  □  □a0684314290 (electronic
book)□z0684312573□  □a(MiFhGG)0684314290□  □aDLC□cDLC□dMiFhGG□00□aE
185□b.S797 2003□1 □aStepto, Gabriel.□14□aThe African American years□h[electronic
resource] /□cGabriel Burns Stepto.□  □aNew York, N.Y. :□bCharles Scribner's Sons,□cc2003.□
□aix, 475 p. :□bill., maps, photos. ;□c29 cm.□ 0□aChronologies of American history and
experience□  □aIncludes bibliographical references and indexes.□  □aAccess restricted to
subscribing institutions.□  □aElectronic reproduction.□bFarmington Hills,
Mi:□cGale,□d2003.□nAvailable via World Wide
Web.□99□anew□d20080123□99□aauth□d20080123□ 0□aAfrican Americans□xHistory.□
0□aAfrican Americans□xHistory□vChronology.□2 □aThomson Gale (Firm)□0 □aGale virtual
reference library.□40□uhttp://lib-
proxy.nd.edu/login?url=http://galenet.galegroup.com/servlet/eBooks?ste=22&docNum=CX340919
9999&q=nd_ref□yClick for ONLINE ACCESS□□
```

Analysis Report



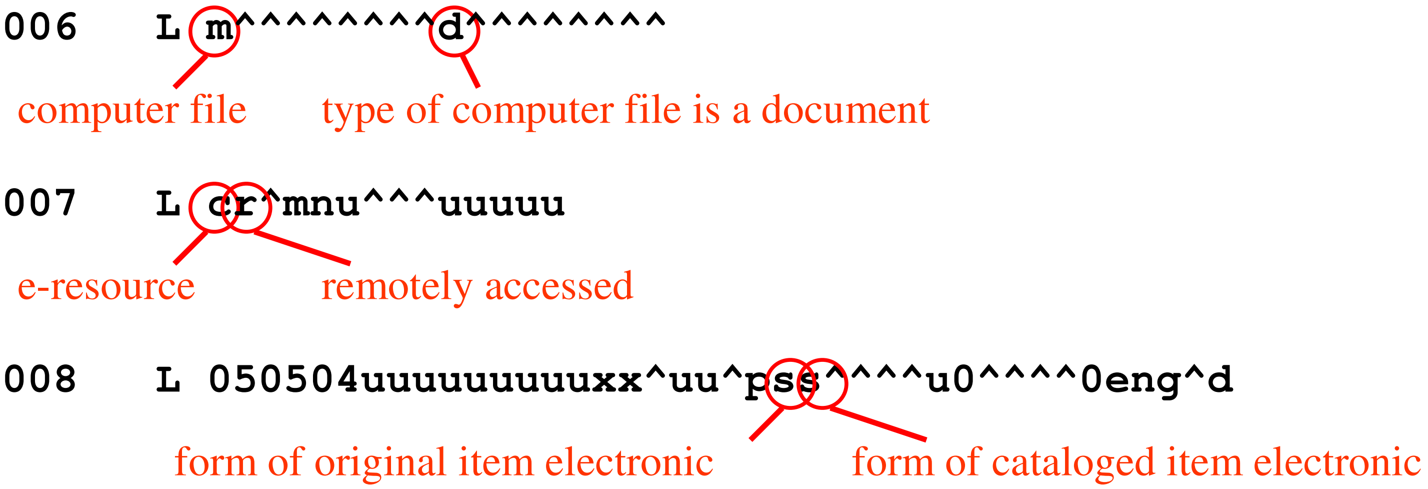
```
Serial, Monograph, Integra
Monograph encoding
  blank: 107
  1: 0
  2: 0
  3: 0
  4: 0
  5: 0
  6: 0
  7: 0
  8: 1
  I: 0
  K: 0
  monograph with 006: 108
  monograph with 006 = m
  monograph with 006 = s:
  monograph with 020: 108
```

```
Field 001
  Total records with an 001 field: 1
  Total records where 001 begins w
  There were no duplicate 001 field

Field 003
  Total records with an 003 field: 1
  Total records where 003 begins w

Field 003
  MiFhGG appears 108 times in fie

Field 006
  Total records with 006: 108
  Total monographs with 006: 108
  Total monographs with 006 with

Field 007
  Total records with 007: 108
  Total records with c & r: 108

Field 020
  Total records with an 020 field: 1
```

```
Field 245
  Total records with a 245 field: 108

Field 260
  Total records with a 260 field: 108

Field 506
  Total records with a 506 field: 0

Field 533
  Total records with a 533 field: 108

Field 534
  Total records with a 534 field: 107

Subjects
  Total records with at least one LCSH heading: 108

Field 856
  Total records with one 856 field: 108
  Total records with one 856 field and subfield 3: 0
  Total records with many 856 fields: 0
```

Control Fields

- **Fixed Length Data Fields (006, 007, 008)**
  - coded with “electronic” characteristics, used in indexes and virtual bases
    - 006s, 007s, and 008s are added when not present
    - 006s and 007s are all removed and re-added for ensured proper coding and consistency (006s and 007s are fairly uncommon on print records)
    - 008 is only modified when it is present (only doesn’t appear on non-enhanced records)



```
006    L m^^^^^^^^d^^^^^^^^
```

computer file

type of computer file is a document

```
007    L cr^mnu^^^uuuuu
```

e-resource

remotely accessed

```
008    L 050504uuuuuuuuuxx^uu^pss^^^^u0^^^^0eng^d
```

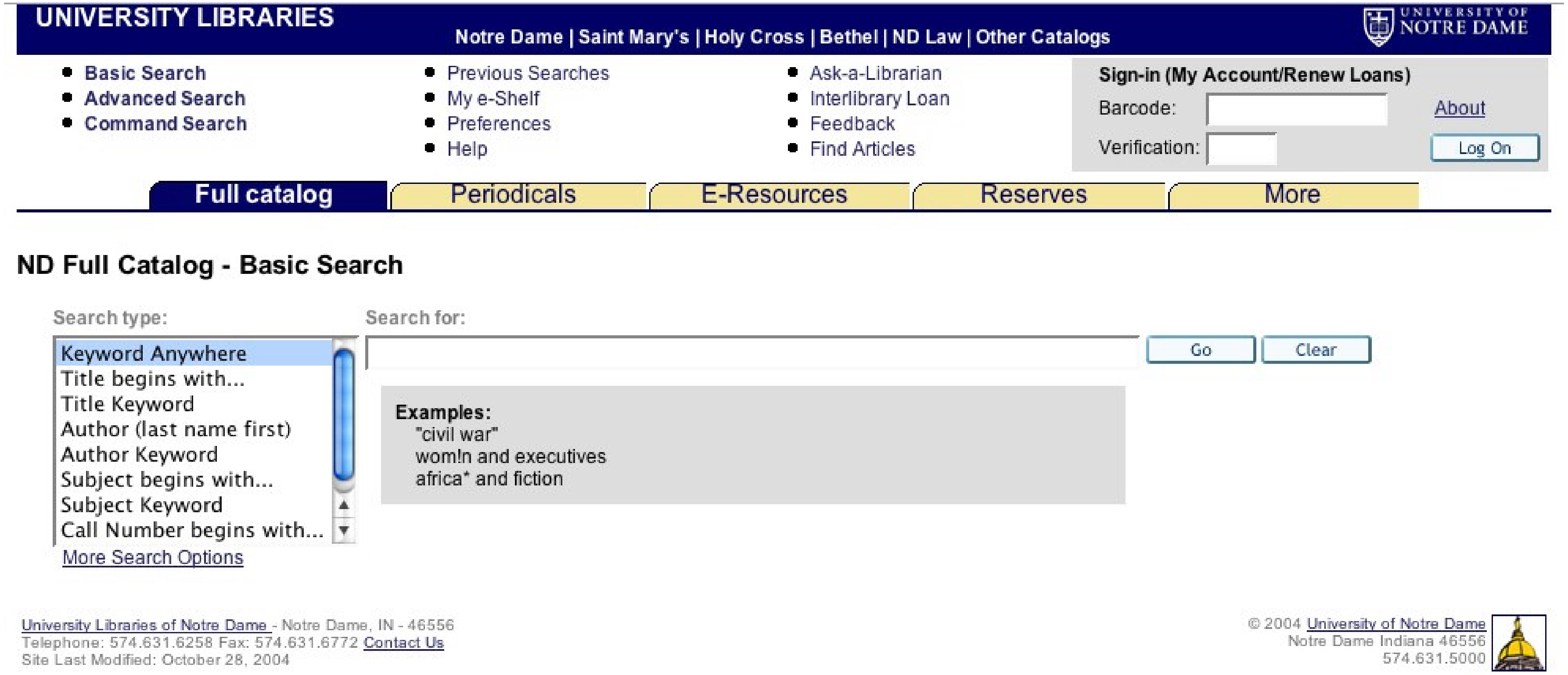
form of original item electronic

form of cataloged item electronic

UNIVERSITY OF NOTRE DAME

Control Fields

Periodicals, E-Resource indexes depend on FMT, 008



UNIVERSITY LIBRARIES

Notre Dame | Saint Mary's | Holy Cross | Bethel | ND Law | Other Catalogs

UNIVERSITY OF NOTRE DAME

- Basic Search
- Advanced Search
- Command Search

- Previous Searches
- My e-Shelf
- Preferences
- Help

- Ask-a-Librarian
- Interlibrary Loan
- Feedback
- Find Articles

Sign-in (My Account/Renew Loans)

Barcode:

Verification:

About

Log On

ND Full Catalog - Basic Search

Search type:

Keyword Anywhere
Title begins with...
Title Keyword
Author (last name first)
Author Keyword
Subject begins with...
Subject Keyword
Call Number begins with...

More Search Options

Search for:

Examples:
"civil war"
wom!n and executives
africa* and fiction

University Libraries of Notre Dame - Notre Dame, IN - 46556
Telephone: 574.631.6258 Fax: 574.631.6772 Contact Us
Site Last Modified: October 28, 2004

© 2004 University of Notre Dame
Notre Dame Indiana 46556
574.631.5000

Record Changes

```
000016212 FMT   L SE
000016212 LDR   L -----nas--2200289-a-4500
000016212 006   L mn^88026116^^^^^^^
000016212 003   L SFXmnu^^^uuuuu
000016212 008   L 880803c19879999mx-qr-pss-----0----0spa--
000016212 010   L $$asn-88026116-
000016212 022   L $$y0187-5337
000016212 035   L $$a(OCoLC)18307370
000016212 035   L $$a(SFX)991042748139142
000016212 040   L $$aDNLM$$cDNLM
000016212 0410  L $$aspa$$beng
000016212 042   L $$alcd
000016212 06000 L $$aW1$$bPE788HJ
000016212 1300  L $$aPerinatologÃ-a y reproducciÃ³n humana  (Online)
000016212 24500 L $$aPerinatologÃ-a y reproducciÃ³n humana$$h[electronic resource]
000016212 260   L $$aMexico :$$bInstituto Nacional de PerinatologÃ-a,
000016212 300   L $$av. :$$bill., ports.
000016212 310   L $$aQuarterly
000016212 3621  L $$aBegan with: Vol. 1, no. 1 (enero-marzo 1987).
000016212 538   L $$aMode of access: World Wide Web.no. 4 (oct.-dic. 1987); title from cover.
000016212 546   L $$aArticles in Spanish; summaries in English and Spanish.
000016212 590   L $$aSFX MARCIt Record
000016212 650 2 L $$aPerinatology$$vPeriodicals.
000016212 650 2 L $$aReproduction$$vPeriodicals.
000016212 7102  L $$aInstituto Nacional de PerinatologÃ-a (Mexico)
000016212 856   L $$uhttp://findtext.library.nd.edu:8889/ndu_local?sid=sfx:opac_856&issn=0187-5337
&pid=serviceType=getFullTxt&genre=journal
000016212 866   L $$xScielo:Full Text$$a Availability: from 2002 volume 16 issue 1 to 2002 volume 16 issue 4
000016212 CAT   L $$c20040616$$lJNL99$$h1735
000016212 SRC   L $$aCONSER
```

MARCIt Non-enhanced Record (Before)

Journal: Archipielago

```
000016405 LDR   L -----nas-a22-----z--4500
000016405 022   L $$a1402-3357
000016405 035   L $$a(SFX)1000000000018746
000016405 090   L $$a1000000000018746
000016405 245   L $$aArchipielago
000016405 856   L
   $$uhttp://findtext.library.nd.edu:8889/ndu_local?sid=sfx:op
   ac_856&issn=1402-
   3357&pid=serviceType=getFullTxt&genre=journal
000016405 866   L $$xGaleGroup Informe:Full Text$$a
   Availability: from 2002
```

MARCIt Non-enhanced Record (After)

Journal: Archipielago 🟦 Additions 🟨 Modifications ⬜ Deletions

```
000016405 FMT   L SE
000016405 LDR   L ^^^^^nas^a22^^^^^z^^4500
000016405 006   L m^^^^^^^^d^^^^^^^^
000016405 007   L cr^mnu^^^uuuuu
000016405 008   L 050504uuuuuuuuuxx^uu^pss^^^^u0^^^^0eng^d
000016405 022   L $$y1402-3357
000016405 035   L $$a(SFX)1000000000018746
000016405 090   L $$a1000000000018746
000016405 1300  L $$aArchipielago (Online)
000016405 24500 L $$aArchipielago$$h[electronic resource].
000016405 538   L $$aMode of access: World Wide Web.
000016405 590   L $$aSFX MARCIt Record
000016405 85640 L
   $$uhttp://findtext.library.nd.edu:8889/ndu_local?sid=sfx:op
   ac_856&pid=serviceType=getFullTxt&issn=1402-
   3357&genre=journal
000016405 866   L $$xGaleGroup Informe:Full Text$$a
   Availability: from 2002
```

UNIVERSITY OF NOTRE DAME

Record Modifications

- **Some modifications complex**

  - removal or additions of only portions of a string or subfield

  - for example, configuring 130 or 240 requires placing word Online within parentheses if they exist, or within its own parens if they don’t:

```
000000003 1300  L $$aPlanning (Chicago, Ill. : 1969)  changed to
000000003 1300  L $$aPlanning (Chicago, Ill. : 1969 : Online)
```

Loading

Using the “Check Input File” (p_manage_36 in Aleph) routine:

- (Adds, Changes, Deletions) - Determines which records already exist in the system, creates two lists – 1) list of new records, 2) list of update records with proper system number
    - requires an index for the search field
    - this is the search we set up in tab_match:

      SFXA match_doc_gen
      TYPE=IND,TAG=035##,CODE=035A,SUBFIELD=a

And just when you thought it was over …

Don’t forget the Authorities

Our authority maintenance reports exploded, the subject report went from 5 to 73 page

- authority flags for items like:

```
$$aAutomobile industry and
trade$$zJapan$$vStatistics$$vPeriodicals
```

- wrote a visual basic script program to modify the reports in Excel and pull out lines that would be ignored, brought reports back to 7 pages

UNIVERSITY OF NOTRE DAME

Phew!

- Tastes great, less filling!
- No wonder we don’t all load these collections.
- There has to be a better way.

Going Forward

Approaches

1. Load the records dirty, only for discovery
2. Modify to cataloging standards
3. ILS as inventory, decoupled discovery

Dirty Loads Appraoch

- Load records with minimal modifications
- Benefits
  - provides basic access points from catalog
  - requires no extra discovery system
- Drawbacks
  - may pollute the swimming pool
  - may not mix with like items

Si Se Puede Approach

- Making adequate changes to records to meet cataloging standards
  - Optional aggregate changes script

- Benefits
  - The catalog contains every item library has access to
  - Doesn’t require decoupled discovery system

- Drawbacks
  - Expensive in time and staff
  - Requires a significant amount of maintenance
  - Time could be invested on what libraries do well – special collections, etc.

ILS as Inventory

- Database of record for physical items
- Appropriate detail level for reporting, decision support
  - augmented with information from discovery environment for NATC, etc
- Access/Collection-level records for digital collections in ILS
- Items in digital collection loaded into discovery system

ILS as Inventory Approach

- Benefits
  - Decoupled discovery system:
    - trend in next generation systems
    - provides more intuitive findability
    - is less fragile
    - has only critical discovery points
    - can be blown away and repopulated
- Drawbacks
  - Requires a decoupled discovery system
  - Reporting on those things outside the ILS requires an added process, may lose some of the specificity

What we are thinking for the way forward

- Assessment of record on access points, not its completeness and well-formedness
- Load records into decoupled discovery system to provide access
- For the most part, these will be static records going in, maybe define a generic input for all batch records

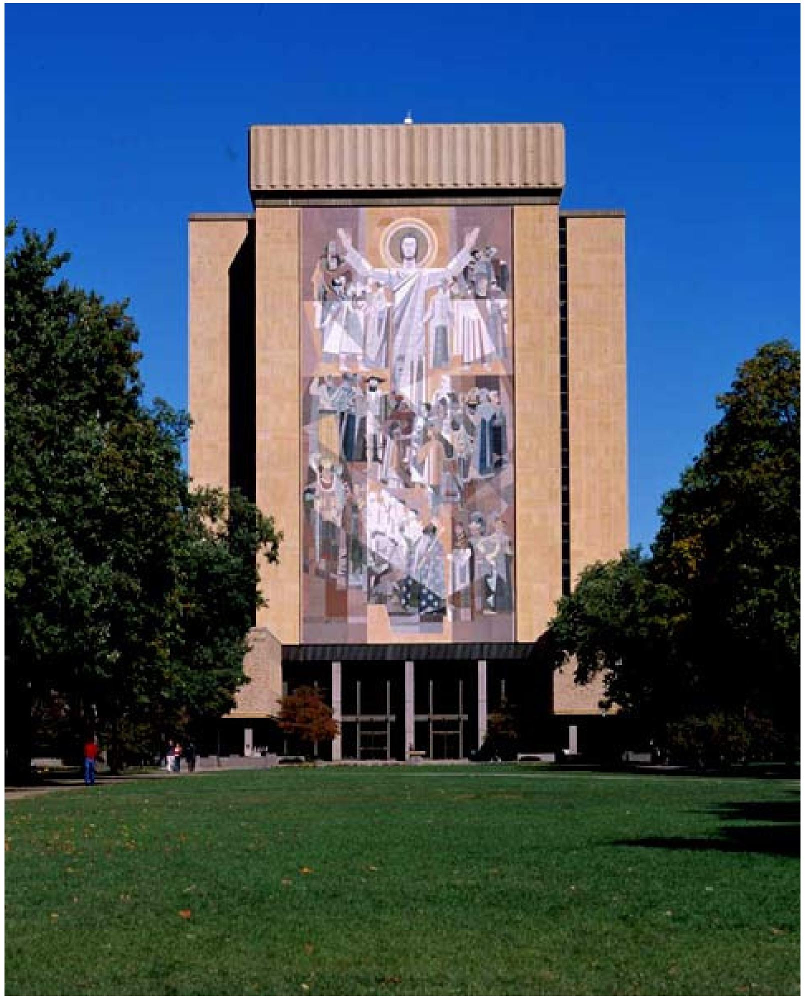
Benefits for User

- Available to user in much faster way – some access is better than NO access
- Catch up on backlog of records, reinvest time on other projects
- Uncovering rich, deep collections that people are unaware of

The URM?

- What does this mean in the context of the URM?
- Relationship to the knowledgebases?
- Localization of data?
- Replication and sharing of data?

Thank You



UNIVERSITY OF NOTRE DAME