Is Archival Description Effective for Discovery?

Jackie Dooley
Consulting Archivist
OCLC Research

SAA Austin
11 August 2009

Overview

- Project objectives
- Problematic characteristics of archival description
- Data mining tool
- Preliminary findings
- Some questions
- Archival discovery issues
    - Discovery = determining existence and location of materials

Project Objectives

- Improve discovery of archival materials in diverse search environments
    - WorldCat, open Web, ArchiveGrid, OPACs …
- Determine system-wide data patterns
    - Inconsistencies, terminology used, busywork …
- Recommend practices to optimize metadata creation
    - Are we including the words that users use in searching?
    - Can we simplify record creation?
    - Are there possibilities for data remediation?
    - What data characteristics are necessary for effective relevance ranking of searches
- Provide system-wide view of underdescribed collections

Problematic Characteristics of Archival Description

- In MARC records ...
  - Vocabulary-rich data elements often lacking
    - MARC 520, 545, 351, 1xx, 6xx, 7xx
  - Inconsistent application of record type (leader 06)
    - Result is inaccurate limiting/faceting of search results
  - Major changes in descriptive rules over time
  - Use of generic titles reduces intelligibility
    - Papers; Records; Letters; Correspondence
- In finding aids ...
  - Hierarchical structure reduces word repetition
  - Weak terms in contents listings: genres, dates ...
  - Abbreviations in lieu of name repetition
  - Brief contents listings (e.g., Correspondence A-Z) in lieu of full enumeration
  - Lack of authority control of names and subjects

Data Mining Tool

- Data population = one million archival MARC records (and counting)
  - Harvested quarterly to build ArchiveGrid
- We can ...
  - Count occurrences of tag groups, fields, subfields
  - Construct complex queries using all Boolean operators
  - Display the content of selected fields and subfields
  - Select randomized query results for full-record analysis
  - Graph usage patterns within and across institutions
- 50,000 finding aids harvested for future study
- OCLC Research colleagues studying the tool's extensibility to other data sets

Preliminary Findings

Demographics

- 93% are held by U.S. institutions

Description

- 36% are minimal-level records
- 57% indicate which cataloging rules were used
- 72% include scope & content note
- 36% include biographical/historical note
- 12% have access/restrictions note
- 23% indicate ownership/acquisition source

Preliminary Findings, cont.

Access points

- 22% have minimal titles (Papers; Records)
- 86% have a principal creator (1xx)
  - 58% are personal names (100)
  - 28% are corporate names (110)
- 33% have personal name added entries (700)
  - 15% have only one occurrence
  - Records exist with up to 466 occurrences
- 11% have corporate name added entries (710)
  - 8% have only one occurrence
  - Records exist with up to 206 occurrences
  - 5% include an organizational subunit
- 48% have genre/form added entries (655)
- Occupation (656, 8%) and function (657, 3%) access points are little-used

Some Questions

- For which data elements is consistency important?
- Should we alter our algorithms for the record type values that constitute “archival material”?
  - p (mixed), t (textual manuscript), k (graphic) ...
  - Optimize for recall or precision?
- What is the nature of the 14% lacking 1xx? Can 1xx’s be derived from other data?
- Can 1xx + 245 be combined to create meaningful titles?
- Is understanding of rules for establishing corporate names (10% have 110$b or 710$b) inadequate? Or do few MARC records for governmental or university records exist?
- Do minimal records include enough data to be discoverable?

Questions, cont.

- Can extent (300) be normalized to improve readability? To prioritize large and therefore “important” collections for discovery purposes?
- Do 520s (scope/content) generally contain high-value words?
- Would it be useful to derive keywords and names from descriptions and normalize as access points?
- Is 7xx (43%) underutilized for lack of standard archival approach for differentiating 6xx vs. 7xx?
- Are genre/form access points (655, 48%) more valued than some believe? If so, are usage guidelines necessary?
- Can 852s (37%) be generated for display of name and location of holding institution
- What % have links to finding aids? Digital objects?

Archival Discovery Issues

- Some sites not yet exposed to search engines
- Recent user studies report says it’s all about ...
  - “Aboutness,” proper names, high-value keywords
    - (Schaffner, *The Metadata* is *the interface* ...)
- Minimal processing = minimal description
  - ... and increasingly this is all we can afford
- Intelligibility of retrieved data
  - Obscure terminology, varying levels of description, hierarchical data, lack of attached digital images
- Usability studies (*ArchiveGrid*, 1990s):
  - Limited search functionality due to data limitations; use of “landing pages” to compensate
- Forthcoming analysis of query logs
  - Proper names (people, places) keywords, title phrases

We Want to Know …

- What questions would you ask of this data?

- What do you see as the chief limitations of archival descriptions for effective discovery?

dooleyj@oclc.org