The Problem With Duplicates

Esmé Cowles
UCSD Libraries, UCAI Project

The Problem

- Large numbers of metadata records from multiple sources
- Significant duplication
  - Between institutions
  - Within institutions

Complicating Factors

- Lack of identifiers
- Inconsistent descriptive practice
- Large scale
- Missing data

Missing Data

- Many records are missing data in key fields
- Reasons for missing data
  - Inapplicable
  - Incomplete description
  - Mapping problems

Clustering

- Similar techniques
- Different assumptions
  - Assumes there are no duplicates
  - Similarity measures that produce a scale of values

Latent Semantic Indexing

- One of several statistical approaches to clustering text
- Requires large amounts of text
- Metadata records are sparse

Normalization

- Removing unimportant dissimilarities
- Case, punctuation, endings, function words, etc.
- Number parsing/formatting
  - Especially for dates

Complex Decisions

- Different strategies for different kinds of records
  - Pre-composed categories
  - Data-driven
- Assign scores to different conditions

Work Around Missing Data

- Missing data in some fields is significant
- Compare only fields that are populated in both records
- Prevent many sparse records from grouping together

Controlled Vocabularies

- Bridge gaps between differing descriptive practices
  - Names
  - Locations
  - Repositories
- Published and ad-hoc vocabularies

Human Intervention

- Ad-hoc vocabularies
- Manually combine or separate records
- In addition to or instead of automatic processes

What Now?

- Standards, standards, standards
  - Controlled vocabularies
  - Format standards
  - Content standards
- Promote identifiers

Contact Info

escowles@ucsd.edu
http://gort.ucsd.edu/ucai/