Assessing Copyright Status –Copyright Renewals Database

7 November 07

What is it?

- Full-text database of the U.S. Copyright Office's published renewal records for books published between 1923-1963

  - Renewals required after 28 years, so renewals from 1950-1992 are relevant

- http://collections.stanford.edu/copyrightrenewals

- Books only – serials, dramas, music etc not included

- Does not track transfers

- Funded by Mellon Foundation

Source Data

- 1950-1977 data extracted from published Catalog of Copyright Entries
  - Project Gutenberg transcripts
  - Edits to Copyright Office internal records are not incorporated in the database
- Electronic records from 1978 forward extracted from Copyright Office online database
  - Also with assistance of Project Gutenberg
  - These records are now more readily available: http://rss.resource.org/

Why?

- Why take on the project?
  - Required input for a copyright analysis system
- Why those years?
  - 1923-1963 renewal registration was required
    - Copyright extended on renewed works only
    - Large numbers of public domain works from this period
- Why is this effort necessary?
  - Access
    - No electronic records pre-'78
    - Multiple searches required

Technology

- Fielded text
  - Much of the text processing was outsourced and managed manually
- Lucene interface
- Data file available on request

Confirming Accuracy - Methodology

- ☐ Two rounds of testing with updates between
- ☐ In each round, ~500 books selected randomly from the Stanford catalog
  - ■ Representative of a research library collection
  - ■ Access to the physical work if required
- ☐ In each round, all 500 books checked manually by a trained student searcher
- ☐ In first round, a subset of 102 books sent to the Copyright Office for searching
  - ■ $100 per hour
- ☐ Final report and full spreadsheets for both rounds of testing on the website

Confirming Accuracy – Round 1

- ☐ Manual search – 545 books searched
  - ■ Initially found 45 discrepancies!
  - ■ 40 = search method, edition or classification issues - rubric reviewed but challenging to correct
  - ■ 1 = error in CCE data
  - ■ 2 = issues with Stanford's import process - corrected
  - ■ 2 = errors in Lesk's parsing of post-1978 data - reparsed

- ☐ Copyright Office – 102 books searched
  - ■ 8 discrepancies – 6 of those parallel manual search
  - ■ Remaining 2 are edition or title problems

Confirming Accuracy – Round 2

- Manual search only – 500 items
  - 26 discrepancies
  - 24 = methodology, edition, or class issues
  - 2 = CRD errors – continuing to investigate source

- Accuracy rate >99%
  - 4 errors in 545
  - 2 errors in 500

- Accuracy does not ensure search success
- Continue to refine database

Lessons Learned

- ~30% of items searched were found to be renewed
  - Earlier estimates were as low as 7%
- Copyright Office data is challenging
  - Formats change year to year
  - No distinct fields in many records – concatenated title and author in some cases
  - Typical bibliographic concerns without as much structure
  - No unique identifiers for this time period
- Errors in the Copyright Office data
  - Patent data in the CCE
  - Missing registration numbers & dates

What's next?

- Incorporate CRD data into a larger system of copyright status analysis
  - Registry files for copyright evidence
  - Measures of certainty
- Continue to refine the data
  - Wiki-type commenting?

Stanford University
LIBRARIES & ACADEMIC INFORMATION RESOURCES

Questions?

Mimi Calter

mcalter@stanford.edu