Capitalising on Metadata

Tool development plans

IASSIST 2007

Chuck Humphrey
University of Alberta
Outline

- Three problems needing metadata solutions
- Three metadata projects
- Two meetings focused on metadata
- Two phases to implement solutions
- The Research Data Centre (RDC) Life Cycle
Three Metadata Problems

The Academic Directors of the RDC Network sought to solve three problems:

1. How to solve the unusable state of the data documentation in the RDC’s, which can be best characterised as huge PDF files repeated for every cycle of a survey. The scale of the data documentation makes a print-based format impractical.

2. How to capture and make available user knowledge about the data acquired from working with the data. How to build a user knowledge database about the data.

3. How to discover related topics across cycles within or across surveys. How to discover comparable variables across surveys more easily.

Proposed to the Canadian Foundation for Innovation (CFI) in December 2005
Project 1: Creating Metadata

- Create DDI-compliant metadata for the existing 50 survey titles in the RDCs now and another 50 titles anticipated over the next four years.

- This project involves separate metadata production and software development tasks. The metadata will be generated through the production task; tools supporting the migration of DDI 1/2 to DDI 3 will be a contract software development task.
Project 2: Capturing User Feedback

- The concept of user feedback fits within the life cycle model of research and needs to be described in the metadata.
- This project will involve two tasks. The first will be to specify the elements of user feedback and building them into DDI 3. The second will be to develop simple tools that are not intrusive for capturing and tagging user feedback.
Project 3: Discovering Comparability

- Enriched metadata will facilitate the discovery of related variables across cycles within or across surveys.

- This project will develop metadata analysis tools that allow researchers to maintain their focus on the conceptual nature of their research questions while letting the tools explore the operational details of working with the data.

Funded by CFI and announced on November 27, 2006
Montreal Metadata Meeting

- Nineteen participants representing organizations interested or involved in metadata tools development met in Montreal on January 25-26, 2007.
  - Eight participants from outside Canada, including Germany, Norway, Switzerland, UK and US.
  - Five from the RDC Network.
  - Six from elsewhere in Canada, including CISTI, HRSDC, Institut de la statistique Québec, Social Sciences Network Data Services (Western University) and Statistics Canada.
Montreal Metadata Meeting

- Maximizing Our Returns from Social Science Data: the case for Data Documentation Initiative 3.0
  - A report prepared by Raymond Currie and Chuck Humphrey following the meeting. Circulated to participants in February 2007.
  - Provides responses to the questions:
    - Why should we invest in metadata?
    - How might we collaborate in developing metadata tools?
  - Gives a summary of discussions about the principles underlying such collaborations.
Metadata RFP Consultation

Three projects organized in two phases:

- After reviewing the details of the three metadata projects, they were organized into two general development phases. Phase one will build foundational architecture for metadata tool development and tools to facilitate the migration from DDI 1.0/2.1 to DDI 3.0.

Phase Two:

- The specifications for phase two will be developed in a subsequent consultation occurring approximately six months after the March 2007 Toronto meeting. The tools in this phase will support mining the metadata for concepts and exploring the metadata for comparable data and variables.
The RDC Research Life Cycle

The RDC Research Life Cycle includes the following stages:

1. Project Application
2. Project Approval
3. Project Creation
4. Access to Data
5. Generate Analysis Files
6. Research Communications
7. Output Disclosure Analysis

These stages represent the life cycle of a research project at the RDC.
Three RDC Processes Requiring Information Management

Managing RDC Projects
- Application processes
- Contract process
- Project account process

Managing Data
- Master files
- Work/analysis files
- File system backups

Managing Research Outputs
- Disclosure process
- Publications
- Knowledge management
Information Needing Management

- Each process generates information:
  - Managing RDC projects: receiving project applications, identifying researchers, conducting peer-reviews, conducting security approval, signing contracts, conducting orientations, assigning project numbers, enabling security access and granting LAN accounts and file space.
  - Managing data: handling master files and their documentation, creating working or analysis files from master files (including syntax files) and backing up the file workspace for projects.
  - Managing research outputs: conducting disclosure analysis, identifying research communications based on RDC research, capturing research outputs and organizing and communicating newly produced knowledge.
The RDC Research Life Cycle

Managing Data Stages

- Project Application
- Project Approval
- Project Creation
- Access to Data
- Generate Analysis Files
- Output Disclosure Analysis
- Research Communications

The RDC Research Life Cycle
Metadata in the Data Stages

Access to master files

Statistics Canada Master Files
- Generate DDI metadata to 1.0/2.0 standards in a retrospective conversion project
- Develop tools to convert DDI 1.0/2.0 to DDI 3.0 and incorporate the Questionnaire module

Repurpose master files

Analysis through Multiple Working Files
- Subset
- Recode
- Compute
- Merge

Multiple Versions
Analysis through Multiple Working Files

Access to master files

Repurpose master files

Multiple Versions

Generate DDI metadata for working files using tools that read statistical system and syntax/log files. These metadata files can be used to document the working files, to produce products from the metadata (e.g., a codebook listing or Powerpoint slides), to be linked to research communications and to recreate the working file from its master data file.

A validation tool will compare a working file against a virtually generated working file.
Metadata in the Data Stages

Analysis through Working Files

Output from Working Files
- Disclosure Analysis
  - Tables
  - Reports
  - Supporting metadata

Research Communications
- Website
- Journals
- Conferences
- Repository

Working Data Files Metadata
Phase Two of the Metadata Project

- The RFP describing the tools to mine and to identify comparable data from the master metadata files will be developed in approximately six months.
- This phase of software development will exploit the architecture built during phase one as well as the DDI metadata for the master microdata files.
Next Steps

- Call for bids on the first phase.
- Staff the Metadata Project Manager position.
- Meet to draft the RFP for the second phase.
- Call for bids on the second phase.