The Evolving Best Practices of DDI in Canada

Jane Fry & Chantal Ripp
IASSIST
June 3, 2015
Agenda

- Background
  - DLI
  - <odesi>
- Our partnership
- DDI
- Best Practices
- The current situation
- Suggestions for the future
In 1996, the Data Liberation Initiative (DLI) became a 5 year pilot project.

Today 78 academic institutions across Canada are partners in the DLI.

Provides access to
  - all Public Use Microdata files (PUMFs)
  - a special collection of aggregate data tables under the DLI License Agreement.
DLI background (cont’d)

- Services include
  - dedicated research assistance
  - support on how to use the Microdata files
  - tools to improve student and faculty access to the data

- In response to demands from our partners, the program adopted the DDI standard for its collection of PUMF files

- In March 2004 acquired the NESSTAR system
Ontario Data Documentation, Extraction Service and Infrastructure

- Established in 2007
- In response to the need for greater data access in Ontario universities

Digital repository for social science data

- Includes microdata, aggregate data and synthetic data files

Part of OCUL (Ontario Council of University Libraries)

- Membership includes 21 university libraries
- Also provides access to other libraries for a fee
<odesi> background (cont’d)

- Nesstar interface used
- DDI standard used
  - Data Documentation Initiative
  - Metadata specification for the social sciences
- Major access point for data in Ontario universities
  - No need of knowledge of a statistical package
  - Access to all associated documentation for the different surveys
- Why <odesi>?
- Role of libraries
DLI and <odesi>: the partnership

- How it came about
  - how it grew

- Didn’t want to re-invent the wheel

- Initial collaboration was between <odesi> and StatCan
  - Carleton U, Queen’s, Guelph and University of Ottawa also involved

- The partnership
  - has always existed informally
Brief Background of DDI

- What is DDI?
  - Data Documentation Initiative

- An international specification
  - Not yet a formal ISO standard

- Goal
  - Formats documentation for social science data files
    - More useful than a word or text file
  - Supports the entire research data lifecycle
Facilitates the creation of metadata

Expressed in XML

XML Schema

- A way of tagging text for meaning, not appearance

Defines

- Which tags are available
- The order the tags will appear in a document
- Whether the tags are required or optional
- Whether the tags are repeatable or not

Different streams of DDI

- DDI Codebook or DDI C or DDI 2.x
- DDI Lifecycle or DDI L or DDI 3.x
- DDI 4 – Coming soon!
Brief DDI Background (cont’d)

- Creates a standard format
  - Used to mark up survey information
  - Meaningful and consistent
  - Metadata is both human and machine readable

- Gives codebook level details such as
  - dataset contents, variable labels, summary Statistics and frequencies
  - Also question text for each variable
    
    *as long as it was entered when marking up the document*
DDI at DLI

- Use DDI Codebook
  - Nesstar is the interface used to interpret the XML
- Labour intensive
  - Marking-up
- Staff turnover
  - DDI and Nesstar not always intuitive
    - Proper training needed
- Need experts in the field to contact
- Development of network of DDI experts
DDI in Academia

- Different universities were doing mark-up at their institutions
  - Did not really know what was happening elsewhere

- Resources
  - Supervisor/trainer
  - Money to hire students

- Many of the same issues as DLI
  - Labour intensive
  - Staff turnover
  - Who to turn to for help

- Bilingual issues
Best practices

- Historical
- Essential
  - When exchanging files
  - Markup is done in geographically diverse locations
    - Consistency needed
- Training documents
  - How to do Markup
- Another reason
- Resources sharing
Best practices (cont’d)

- Quality Control
  - Mark-up is done at different universities
  - How to ensure that BPD is being followed
  - Who does Quality Control (QC)
  - Needed to come up with a list of files being worked on, by whom, …
  - Scholars Portal maintains the list
    - We edit it as necessary
- Main pages
  - QC Activity - microdata files (one for English, one for French)
  - New MarkUp activity (one for English, one for French)
  - QC Activity – Aggregate data files
Example:
1.1.1.5 <IDNo> Identification Number

*Description:* Unique string or number (producer's or archive's number) for the marked-up document.

*Note 1:* This ID number is the same for the document description and the study description, i.e. 1.1.1.5 is the same as 2.1.1.5.

*Note 2:* For Statistics Canada surveys, the catalogue number refers to the microdata file.

*Formatting Notes:*
- Language: E = English, F = French
- Year: yyyy or yyyy-mm-dd
- Lowercase: should be used for everything except the catalogue number and the language abbreviation
- Surveys with cycle numbers and sub-numbers: use a dash between the numbers and not a period, e.g. for cycle 2.1 the format would be c2-1
- ICPSR surveys: use their ID# and add the short form if there is a subset, e.g. icpsr9721im

This is the format to be used: acronym_CatalogueNumber_language_year_subset
Best practices (cont’d)

- Example 1: `<titl>Sun Exposure Survey, 1996 [Canada]</titl>`
  `<IDNo>ses_82M0019_E_1996</IDNo>`

- Example 2: `<titl>General Social Survey, 2005 [Canada]: Cycle 19, Time Use, Main File</titl>`
  `<IDNo>gss_12M0019_E_2005_c19_main-file</IDNo>`

- Example 3: `<titl>Canadian Tobacco Use Monitoring Survey, 2004: Cycle 1, Household File</titl>`
  `<IDNo>ctums_82M0020_E_2004_c1_household-file</IDNo>`

- Example 4: `<titl>Canadian Tobacco Use Monitoring Survey, 2004: Cycle 1, Household File</titl>`
  `<IDNo>cchs_82M0013_E_2005_c3-1_health-sample</IDNo>`

- Example 5: `<titl>Canadian Gallup Poll, May 1949, #186</titl>`
  `<IDNo>cipo_186_E_1949-05</IDNo>`

- Example 6: `<titl>Descriptors and Measurements of the Height of Runaway Slaves and Indentured Servants in the United States: Runaway Indentured Servant Height Measurements, 1700-1850</titl>`
  `<IDNo>9721im</IDNo>`
Currently ...

- Formal review of BDP
- Ad hoc updates
  - e.g. old ID Number: CatalogueNumber_language_year_subset
  - e.g. new ID Number: CatalogueNumber-language-year-subset
- Creating and sharing of new best practices
  - Aggregate data markup
  - Creating Variable Groups
  - Creation of training videos
- Meetings
  - Carleton U and Queen’s
  - Scholar’s Portal MarkIt! team
  - Recent meeting of partners (DLI and <odesi>)
Suggestions for the future

- Continue with the present collaboration
- Set up a working group
  - Working Group on File Enhancement
- Expand collaboration with internal and external stakeholders

Question to ponder
- Who will be the overseer of DDI and Best Practices?
  - Public services?
  - Academia?
In sum

- We discussed
  - The background of DLI and <odesi>
  - How our partnership has worked over the years
  - Our background with DDI
  - The current and evolving situation with Best Practices
  - Suggestions for the future

- DLI and <odesi>
  - A valuable partnership for sustainability
References

- **Best Practices Documents:**
  - <odesi>: [http://www.library.carleton.ca/help/nesstar-how-mark-nesstar](http://www.library.carleton.ca/help/nesstar-how-mark-nesstar)
  - Statistics Canada - for a copy, contact: dli-idd@statcan.gc.ca
  - Carleton Data CentreYouTube: [https://www.youtube.com/user/MacOdrumLibrary/playlists](https://www.youtube.com/user/MacOdrumLibrary/playlists)
  - Marking-up in Nesstar videos: [http://www.library.carleton.ca/help/nesstar-how-mark-nesstar](http://www.library.carleton.ca/help/nesstar-how-mark-nesstar)

- **Presentations:**
  - Fry, J. “Discover the power of DDI metadata”. Workshop at NADDI April 2014. [http://summit.sfu.ca/item/13941](http://summit.sfu.ca/item/13941)
Jane Fry
Data Specialist
MacOdrum Library
Carleton University
613.520.2600 x1121
jane.fry@Carleton.ca

Chantal Ripp
Unit Head
Data Liberation Initiative (DLI)
Statistics Canada
613.951.3817
Chantal.Ripp@statcan.gc.ca
Questions?