Metadata driven tools developed for the Canadian Research Data Centre Network (CRDCN)

IASSIST 2013 - May 29, 2013

Donna Dosman and Pamela Moren, Statistics Canada
Chuck Humphrey, University of Alberta
Pascal Heus and J.Gager, Metadata Technology
Background

Stakeholders:
- Canadian Research Data Centre Network (CRDCN)
  - Network of 27 Canadian Universities
- Canadian Foundation for Innovation (CFI)
- Statistics Canada – Microdata Access Division

Roles:
- CRDCN obtained CFI grant, planned project, hired contractor
- Statistics Canada managed project
Goal of project

• Conduct a retrospective conversion of survey documentation into DDI-Lifecycle standard

• Develop metadata tools for data management and research purposes that leverage DDI-Lifecycle metadata
Coding DDI Metadata

• Original list in 2005 consisted of 182 cycles representing 44 surveys

• Manual coding of metadata by 2-11 coders (2009-2013) using Nesstar Publisher software

• Automatic coding of metadata of study description metadata using the IMDB-DDI conversion tool

• Automatic DDI conversion from DDI-C to DDI-L XML using the DDI-C to DDI-L upgrade tool
Tools for DDI conversion

Part of the Research Data Centre Metadata Project:
- DDI-C to DDI-L upgrade tool
- Colectica software

Separate StatCan DDI initiative:
- IMDB-DDI conversion tool
Abstract

The Aboriginal Peoples Survey is a national survey of Aboriginal peoples (First Nations peoples living off-reserve, Métis and Inuit) living in urban, rural and northern locations throughout Canada. The survey provides valuable data on the social and economic conditions of Aboriginal children and youth (6-14 years) and Aboriginal people (15 years and over).

The Aboriginal Peoples Survey was developed and implemented in partnership with the following National Aboriginal Organizations:
- Congress of Aboriginal Peoples (CAP)
- Inuit Tapiriit Kanatami (ITK)

Kind of Data

Sample survey data [ssd]

Unit of Analysis

Person
Nesstar Publisher
Labour Force Survey - November 2000

“Labour Force Survey - November 2000”.

Abstract

The Labour Force Survey provides estimates of employment and unemployment which are among the most timely and important measures of performance of the Canadian economy. With the release of the survey results only 13 days after the completion of data collection, the LFS estimates are the first of the major monthly economic data series to be released.

The Canadian Labour Force Survey was developed following the Second World War to satisfy a need for reliable and timely data on the labour market. Information was urgently required on the massive labour market changes involved in the transition from a war to a peace-time economy. The main objective of the LFS is to divide the working-age population into three mutually exclusive classifications - employed, unemployed, and not in the labour force - and to provide descriptive and explanatory data on each of
### Labour Force Survey - October 2000

**Showing 1 to 100 of 187 entries**

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Data Type</th>
<th>Format</th>
<th>Notes</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERTAB</td>
<td>Economic regions</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>Economy</td>
</tr>
<tr>
<td>CMATAB</td>
<td>Census metropolitan ar...</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>Census</td>
</tr>
<tr>
<td>UIRTAB</td>
<td>Unemployment insur...</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>Unempl</td>
</tr>
<tr>
<td>PROV1</td>
<td>1-digit province code</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>1-digit</td>
</tr>
<tr>
<td>PSEUDOU1</td>
<td>Pseudo Unemployment...</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>Pseudo</td>
</tr>
<tr>
<td>FRAME</td>
<td>Geographic/regional str...</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>Large</td>
</tr>
<tr>
<td>STRA1RAM</td>
<td>Usually a super-stratum</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>Usually</td>
</tr>
<tr>
<td>TYPF</td>
<td>Sample design type</td>
<td>Code</td>
<td>Code Format</td>
<td>0</td>
<td>Sample</td>
</tr>
</tbody>
</table>
IMDB-DDI
<table>
<thead>
<tr>
<th>SDDS</th>
<th>Title</th>
<th>Dataid</th>
<th>Version</th>
<th>Itemld</th>
<th>Dissemination</th>
<th>DDI-XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>3901</td>
<td>Census of Population</td>
<td>3901</td>
<td>0</td>
<td>1238</td>
<td>2:PUBLIC</td>
<td>By Ref</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[urn:ddi.ca.statcan:StudyUnit.imdb13723.1.0.0]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[urn:ddi.ca.statcan:StudyUnit.imdb13723.2.0.0]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[urn:ddi.ca.statcan:StudyUnit.imdb13723.3.0.0]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[urn:ddi.ca.statcan:StudyUnit.imdb13723.4.0.0]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Catalog/Manager

Select a zip file containing study DDI-L (version 3.1) export from the IMDB.

Browse
Submit Query
Catalog/Manager

CRDCN Master Catalog

Census of Population [hide]
  Census of Population - 1996
  Census of Population - 2001
  Census of Population - 2006

Labour Force Survey [hide]
  Labour Force Survey - October 2000
  Labour Force Survey - November 2000
  Labour Force Survey - December 2000
Outcome for the researcher
<table>
<thead>
<tr>
<th>Study</th>
<th>Name</th>
<th>Label</th>
<th>Type</th>
<th>Format</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Community Health Sun</td>
<td>SAMPLEID</td>
<td>Household identifier</td>
<td>Integer</td>
<td>Numeric</td>
<td>H</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>PERSONID</td>
<td>Person identifier of person selected</td>
<td>Integer</td>
<td>Numeric</td>
<td>P</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOB_PRV</td>
<td>Province</td>
<td>Code</td>
<td>Code Format</td>
<td>P</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOB_HR4</td>
<td>Health region</td>
<td>Code</td>
<td>Code Format</td>
<td>P</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOB_PC</td>
<td>Postal code</td>
<td>Text</td>
<td>Text</td>
<td>P</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOBDEA</td>
<td>Enumeration Area</td>
<td>Integer</td>
<td>Numeric</td>
<td>E</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOBDFED</td>
<td>Federal Electoral District</td>
<td>Code</td>
<td>Code Format</td>
<td>P</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOBDCSD</td>
<td>Census Sub-Division</td>
<td>Integer</td>
<td>Numeric</td>
<td>C</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOBDCD</td>
<td>Census Division</td>
<td>Code</td>
<td>Code Format</td>
<td>C</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOBDMCA</td>
<td>Census Metropolitan Area</td>
<td>Code</td>
<td>Code Format</td>
<td>C</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOBDUR2</td>
<td>Rural/Urban Area - 2 levels</td>
<td>Code</td>
<td>Code Format</td>
<td>R</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>GEOBDURS</td>
<td>Rural/Urban Area - 5 levels</td>
<td>Code</td>
<td>Code Format</td>
<td>R</td>
</tr>
<tr>
<td>Canadian Community Health Sun</td>
<td>SAMB_ACP</td>
<td>Actual collection period</td>
<td>Code</td>
<td>Code Format</td>
<td>A</td>
</tr>
</tbody>
</table>

Variable: GEOBDCSD
Label: Census Sub-Division
Description:
Summary of Tools

For Data Managers:
- DDI-Lifecycle core editor (Metadata Manager)
- RDC Catalog / Manager
- Quality assurance utility
- DDI-C to DDI-L upgrade tool
- Blaise to DDI-L converter
- Colectica RDC Edition software

For Researchers:
- DDI-Lifecycle researcher’s tool (Dataset Builder)
Where we are now

- Coded 243 cycles representing 39 surveys in DDI-C
- Upgraded 223 cycles representing 38 surveys in DDI-L
- Coded over 400,000 variables for both English and French in DDI-C and DDI-L
- DDI-L prototype applications and tools have been deployed into RDC environment
Next Steps

• StatCan metadata working group

• DDI catalog for public-use and master microdata files

• RDC Data Management Tool

• Integrating DDI into StatCan’s generic systems/tools for social and business surveys

• Creation of a new DDI Centre

• Further development on RDC Dataset Builder
Acknowledgements

- Statistics Canada (StatCan)
- Canadian for Innovation (CFI)
- Canadian Research Data Centre Network (CRDCN)
- Metadata Technology North America (MTNA)
- Algenta Technologies
- Breckenhill
Contacts

Pamela Moren: pamela.moren@statcan.gc.ca
Donna Dosman: donna.dosman@statcan.gc.ca
Chuck Humphrey: humphrey@Datalib.library.ualberta.ca
Pascal Heus: pascal.heus@metadatatechnology.com
Jack Gager: j.gager@metadatatechnology.com
Jeremy Iverson: jeremy@colectica.com
Dan Smith: dan@colectica.com