



Metadata driven framework for the Canada Research Data Centre Network

IASSIST 2010 – Session A4: DDI3 Tools

Pascal Heus, Metadata Technology North America

pascal.heus@metadatatechnology.com

<http://www.metadatatechnology.com>



1428 Washington Ave S Ste 203
Minneapolis, MN 55454
USA

<http://www.algenta.com>



305 Breckenridge Cres.
Ottawa, Ontario, K2W 1J3
Canada

<http://www.breckenhill.com>
contactus@breckenhill.com



Lyngveien 15
N-5101 Eidsvågneset
Norway

<http://www.ideas2evidence.com>



200 Prosperity Dr
Knoxville, TN 37923
USA

<http://www.metadatatechnology.com>
mtna@metadatatechnology.com



Background

- Canadian Research Data Centre Network (CRDCN - <http://www.rdc-cdr.ca>) established in 2000/2001
- 24 research data centres located in universities across Canada.
- Secure access for approved researchers to confidential micro-data (primarily Statistics Canada)
- Provides computer resources and technical support for analysis
- One Statistics Canada analyst is present in each centre to assist researchers and to ensure data confidentiality
- All centers were recently connected over a secure Intranet network
 - Can now provide central/virtual access to data and metadata
 - Potential for collaborative research



Project Overview

- Implement a DDI3 driven enterprise platform for data management, discovery, access, and analysis across the entire CRDCN
- DDI 3.0 compatible metadata on over 60 data titles (hundreds of surveys, millions of variables)
- Key Components
 - Back-end data / metadata storage (files, XML)
 - Middle-ware layer: web service oriented architecture
 - Management tools: data / metadata administration
 - Researcher tools: discovery, data customization, analysis, capture of research process / data usage
- Project planned over a two year period
 - Initiated June 2009
- A second phase aiming at the implementation of advanced tools for harmonization/comparability, complex metadata exploration, disclosure control is expected to follow



Project Participating Agencies

- CRDCN / University of Manitoba
 - Project coordination / management
- Canada Foundation for Innovation (CFI)
 - Project funding
- Development partners
 - Breckenhill, Ottawa, Canada
 - Metadata Technology, Knoxville, TN, USA
 - Algenta, Minneapolis, MN, USA
 - Ideas2evidence, Bergen, Norway
- CRDCN Research Metadata Centre
 - Established in Ottawa for the capture of survey data and metadata
- Statistics Canada
 - Provides data and documentation



Products

- Enterprise platform to meet CRDCN requirements
 - Not a generic platform, but with focus on reusability / extension
- Metadata storage (IBM/DB2)
 - Hybrid relational / XML. Free and commercial license available
- Data and file storage (iRODS)
 - Open source virtual file system
- Metadata registry / web services
 - J2EE, Tomcat, Spring
- Desktop products:
 - Common application framework
 - Data/Metadata management suite + Colectica RDC Edition
 - Researcher Suite
 - Based on Eclipse/RCP. Uses BaseX for local metadata storage
- To be released under an open source license



Data/Metadata Flow

- Capture initial data / metadata using DDI2 (IHSN Metadata Editor)
- Use upgrade tools to convert to DDI3 and upload of data and documentation files to data repository (multilingual)
- Use DDI3 Management Suite to enhance metadata, harmonize within study unit (variables, classifications, questions, etc.). Use Colectica RDC Edition for questionnaires
 - Note that data is read only and structure / content cannot be changed
- Control quality and “publish” into master catalog (registry/repository)
- Grant researchers access to relevant catalogs based on project
- Researcher discover data by study, variable, question, concept, etc.
- PI prepares a “virtual dataset” (DDI3) that meets the need of the research topic (subset of variables/observations, recodes, etc)
 - Used to automatically generate ASCII + imports scripts for statistical packages
- Research team describes the “research process” using workflow metadata
- Use various reporting tools to document processes, data/variable usage, etc.



Server side infrastructure

- Technologies
 - J2EE / Java / Spring (security, web services, MVC) / SOA
 - DDI3 Metadata registry for search and retrieval
- Metadata storage: IBM/DB2
 - Free ExpressC or licensed version
 - Solid support for both relational and XML
 - Full text search available
- Data storage: iRODS
 - Virtual file system (abstraction of back en infrastructure, rule engine, backup/mirroring, federation, open source, etc.)
 - Associate file with metadata (i.e. DDI3 URN, Dublin Core, etc.)
- Security
 - Desktop / OS based authentication
 - Integration in CRDCN LDAP: project level access authorization (users have unique account per project)
 - Use WS-Security for desktop application communication
- Virtual server farm (VMWare) and NAS for storage



Common Application Framework

- Components shared by all desktop applications (management, research, ...)
- Provide features such as:
 - Help, multilingual support, automatic updates, preferences, install, libraries
 - Integration in security system
 - Web services
 - Local metadata repository
- Technologies
 - Eclipse Rich Client Platform (desktop application)
 - Spring (injection, MVC)
 - BaseX (local repository)
 - Sferyx as Rich Text Editor (tinyMCE available as well)



Management Suite

- Maintain metadata and data across system
- Access to master survey catalog (based on group membership)
 - All for Metadata Admin, subset for Metadata Operators
- Import from DDI2
- View or check out / check in survey for updates
- Access to custom editors:
 - Study, Variables, Datasets, Classifications, External resources, etc.
- Local save and repository commit
- Integration with Colectica RDC Edition for questionnaires
- Publication workflow
 - Operator submits for approval and administrator approves
 - Subsequent changes require versioning (with some exceptions)



Some Design Techniques

- Extensive use of code injection
 - Editor is a collection of “widgets” described in XML
 - Low level widgets are typically DDI reusable types
 - Allows for different widgets for same type (Citation, dates, etc.)
 - Allows for dynamic interface and reusability outside the project
- Editor synchronization within a Study Unit
 - Sharing same object (bean) and using events
- Implemented generic object (beans) to abstract DDI version (or deal with DDI features/bugs)
- Check in / check out for concurrent editing
 - A local save or registry commit is always at the Study Unit level (for referential integrity!)
- Three level of metadata storage
 - Cache (in memory, very fast), Local (in BaseX, fast), Remote (call the registry, no as fast)
 - Metadata elements are retrieved at the maintainable level on a “as needed basis”



Catalog View

The screenshot displays the Metadata Editing Suite interface. On the left, the 'Catalog Explorer' pane shows a search for 'agri' with results including 'Census of Agriculture 2002' and various 'StudyUnit catalog' entries. A 'Quick Search' box is positioned above the search input. Below the catalog list, an 'Editors' box lists actions like 'Edit StudyUnit', 'Edit Questions', and 'Edit Concepts'. The main 'Overview Editor' pane displays the details for the 'Kauffman Firm Survey: Baseline/First/Second/Third Follow-Up', including its title, author information, abstract, and a list of indicators. A 'Study Overview' box is located at the top right of the main pane.

Quick Search

agri Search

Study Catalogs

- Search Results
 - [en]Census of Agriculture 2002
 - [en]Household Income and Expenditure Sur
- Remote Catalogs
 - Test completeness 2005 title
 - [en]StudyUnit catalog for 'IHSN' collection
 - [en]StudyUnit catalog for 'BLR' collection
 - [en]StudyUnit catalog for 'CMR' collection
 - [en]StudyUnit catalog for 'COD' collection
 - [en]StudyUnit catalog for 'CUB' collection
 - [en]StudyUnit catalog for 'ETH' collection
 - [en]StudyUnit catalog for 'GHA' collection
 - Core Welfare Indicators Questionnaire 20
 - Edit StudyUnit
 - Edit Questions
 - Edit Concepts
 - Edit Variables
 - Edit Documentation
 - Edit Dataset
 - [en]StudyUnit catalog for 'GMB' collection
 - [en]StudyUnit catalog for 'HND' collection
 - Encuesta Permanente de Hogares de Pro
 - [en]StudyUnit catalog for 'IRQ' collection
 - [en]StudyUnit catalog for 'JAM' collection
 - [en]StudyUnit catalog for 'KGZ' collection
 - [en]StudyUnit catalog for 'LBR' collection
 - [en]StudyUnit catalog for 'LKA' collection
 - [en]StudyUnit catalog for 'MAR' collection
 - [en]StudyUnit catalog for 'MDG' collection
 - [en]StudyUnit catalog for 'MNG' collection
 - [en]StudyUnit catalog for 'NER' collection
 - [en]StudyUnit catalog for 'NGA' collection
 - [en]StudyUnit catalog for 'Other' collection

Editors

- Edit StudyUnit
- Edit Questions
- Edit Concepts
- Edit Variables
- Edit Documentation
- Edit Dataset

Study Overview

Kauffman Firm Survey: Baseline/First/Second/Third Follow-Up

"Kauffman Firm Survey: Baseline/First/Second/Third Follow-Up" by Alicia Robb, David DesRoches, 2007-10-12, Copyright: Kauffman Firm Survey ©2008 by Ewing Marion Kauffman Foundation.

Abstract

The stability of the American economy depends on the spirit of entrepreneurship that drives the creation of new businesses, jobs, and innovations. Yet launching and running a business can involve many challenges, and many efforts fail in their early years. These barriers to success threaten our nation's long-term prosperity in an increasingly competitive global economy. A new study aims to help new business owners overcome start-up challenges and build innovative, growing companies.

The Ewing Marion Kauffman Foundation of Kansas City is sponsoring this research to gain a better understanding of how firms grow, strengthen, and mature. Mathematica is conducting a study of new businesses to help the Foundation in its efforts to promote new business development.

The Kauffman Firm Survey (KFS) is the largest longitudinal study of new businesses ever embarked upon. The panel of businesses was created by using a random sample from Dun & Bradstreet's (D&B) database list of new businesses started in 2004, which totaled roughly two hundred fifty-thousand such businesses. The KFS oversampled these businesses based on the intensity of research and development employment in the businesses' primary industries. The KFS sought to create a panel that included new businesses founded by a person or team of people, purchases of existing businesses by a new ownership team, and purchases of franchises. To this end, the KFS excluded D&B records for businesses that were wholly owned subsidiaries of existing businesses, businesses inherited from someone else, and not-for-profit organizations. Also, previous research on new businesses has reported variability in how business founders perceive when their businesses started operations. Therefore, a series of questions were asked of business owners about indicators of business activity and whether these were conducted for the first time in the reference year (2004). These indicators included:

- * Payment of state unemployment (UI) taxes
- * Payment of Federal Insurance Contributions Act (FICA) taxes
- * Presence of a legal status for the business

System Information

Study Editor

The screenshot displays the Metadata Editing Suite interface, specifically the Study Editor for the 'Core Welfare Indicators Questionnaire 2003'. The interface is divided into several panels:

- Catalog Explorer:** Located on the left, it shows a search for 'agri' and a list of search results. The selected item is 'Core Welfare Indicators Questionnaire 2003', with options to 'Edit StudyUnit', 'Edit Questions', 'Edit Concepts', 'Edit Variables', 'Edit Documentation', and 'Edit Dataset'. Below these are various study unit catalogs for different countries and collections.
- Overview Editor:** The main editing area, titled 'Core Welfare Indicators Questionnaire 2003'. It contains several fields and widgets:
 - Citation:** A section with a 'Citation Widget' label. Fields include Title (Core Welfare Indicators Questionnaire 2003), Subtitle, Abbreviation (CWIQ II), Language, and Copyright.
 - Date:** A field containing '11/28/05' with a 'Date Widget (preferences driven)' label and a 'Choose Date' button.
 - Creators:** A field containing 'Ghana Statistical Service' with a 'Creators Widget (comma separated)' label.
 - Contributors:** A section with a 'Contributors Widget (tabular form)' label, including a table with a 'Name' column and a 'Select All / Unselect All' control.
 - Abstract:** A section with a 'Structured String Widget' label and a 'Rich Text Editor' label, featuring a rich text toolbar with options like Bold, Italic, Underline, and text color.

Variable Browser / Editor

The screenshot displays the Metadata Editing Suite interface, which is divided into several functional areas:

- Filters Panel (Left):** Contains sections for 'Filter', 'Files' (with checkboxes for File Filter 0, 1, and 2), 'Type' (with checkboxes for Discrete, Continuous, Weight, Derived, Time, and Geography), 'Question' (with checkboxes for No Question and Has Question), 'Concept' (with checkboxes for No Concept and Has Concept), and 'Universe' (with checkboxes for No Universe and Has Universe). It also includes expandable sections for 'Groups', 'Concepts', and 'Universe'.
- Variable Browser (Center):** A table listing 75 variables. The table has columns for Name, Label, DataType, Classification, Question, Concept, and Universe. The variable 'q0561' is selected and highlighted in blue. A callout box points to this table with the text: 'Variable Browser With support for harmonization, multiple selections and custom column views'.
- Metadata Editor (Bottom):** A form for editing the selected variable 'q0561'. It includes fields for Name, Label, Description, Universe, Concept, Question, and Response Unit, each with an 'Edit' button. There are also checkboxes for 'Is Time', 'Is Geographic', and 'Is Weight'. A 'Data Type' dropdown is set to 'Code'. A callout box points to this editor with the text: 'Variable Editor Additional tabs to show summary statistics, etc.'.
- Overview Editor (Top):** Shows the current project context, including 'Overview Editor', 'CWIQ II', and 'variableEditor'.
- Search and Navigation (Top):** Includes a search bar with the text 'Display options and quick search', and pagination controls showing 'Found 75 variable(s)', 'Page 1 of 2', and 'Show 50'.



Researcher Suite

- Support data discovery across various metadata dimensions (using others as constraint)
 - Variable, study, time/geography, etc.
- Access to all documentation
- Production of virtual datasets
 - Select subset of variables
 - Select subset of cases
 - Simple data transformations (recodes, banding)
 - Retrieve ASCII data + generated import scripts (SPSS, SAS, Stata, etc.)
- Capture of research process
 - Personal and team project log with links to metadata elements
 - Description of analytical process flow
- To be further discussed with researchers /users



DDI Upgrade Tool

- Command line utility driven by a XML configuration file (wrapped in application wizard)
- Currently converts DDI 1.2.2 into DDI 3.1 (2 languages)
- Multi-stages (with info, warning, error)
 - DDI 2 schema, second level and custom validation
 - Availability of external resources
 - DDI 3 upgrade and validation
 - Multilingual merge with cross DDI validation to ensure consistency
 - Upload data and external resource to iRods
- Use code injection to facilitate customization
- Private beta testing over summer
 - Contact us if interested to contribute
- Planned for open source release Oct 2010



Status and next steps

- Basic architecture is in place
- Common application framework completed
- DDI Upgrade tool in closed beta
 - Public release October 2010
- Management suite to be deployed at RMC Ottawa for beta testing this summer
 - Study, Variable, Classification, etc.
 - Public release 4Q 2010
- Researcher Suite development to begin later this year. Release planned in 2011.
- Other activities
 - Support metadata preparation by RMC Ottawa (now fully staffed)
 - Ongoing collaboration with Statistics Canada for extracting public metadata from IMDB with potential conversion to DDI



Congratulations to Raymond Currie

- 2010 recipient of the Lise Manchester Award as Executive Director of the CRDCN
- Recognizes excellence in statistical research
- *“for his leadership role and vision in bringing the network to a high level of excellence in the promotion and use of a broad range of microdata for research work that has influenced the formation of social and health policies in Canada.”*
- Key accomplishments
 - 5-year grant of \$ 1.6 million from SSHRC
 - **4-year award from CFI for Lightpath Intranet and DDI 3.0 metadata for over 60 datasets**
 - Upcoming 3-year research contract or up to \$ 1 million for social policy contract research
 - In this decade, the CRDCN supported over 1200 projects and 2600 research, including 1000 graduate students, which has lead to over 1000 publications
- <http://www.ssc.ca/en/award-winners/award-winners-2010#manchester>





THANK YOU!

Q&A?



1428 Washington Ave S Ste 203
Minneapolis, MN 55454
USA

<http://www.algenta.com>



305 Breckenridge Cres.
Ottawa, Ontario, K2W 1J3
Canada

<http://www.breckenhill.com>
contactus@breckenhill.com



Lyngveien 15
N-5101 Eidsvågneset
Norway

<http://www.ideas2evidence.com>



200 Prosperity Dr
Knoxville, TN 37923
USA

<http://www.metadatatechnology.com>
mtna@metadatatechnology.com

