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

algenta
1428 Washington Ave S Ste 203
Minneapolis, MN 55454
USA
http://www.algenta.com

BRECKENHILL
305 Breckenridge Cres.
Ottawa, Ontario, K2W 1J3
Canada
http://wwww.breckenhill.com
contactus@breckenhill.com

ideas 2 evidence
Lyngveien 15
N-5101 Eidsvågneset
Norway
http://www.ideas2evidence.com

metadata technology
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)

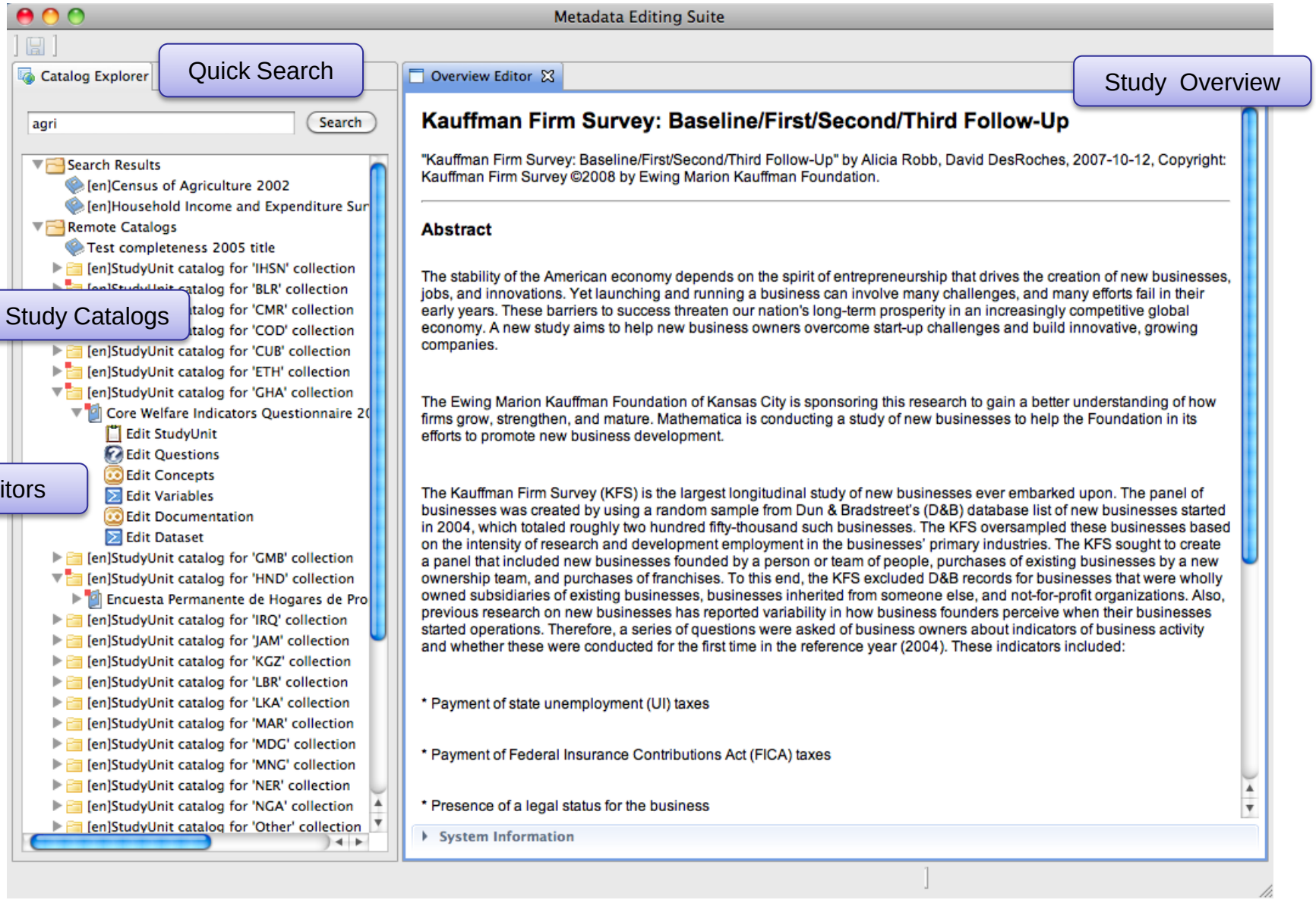
Management Suite

- 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 excpetions)

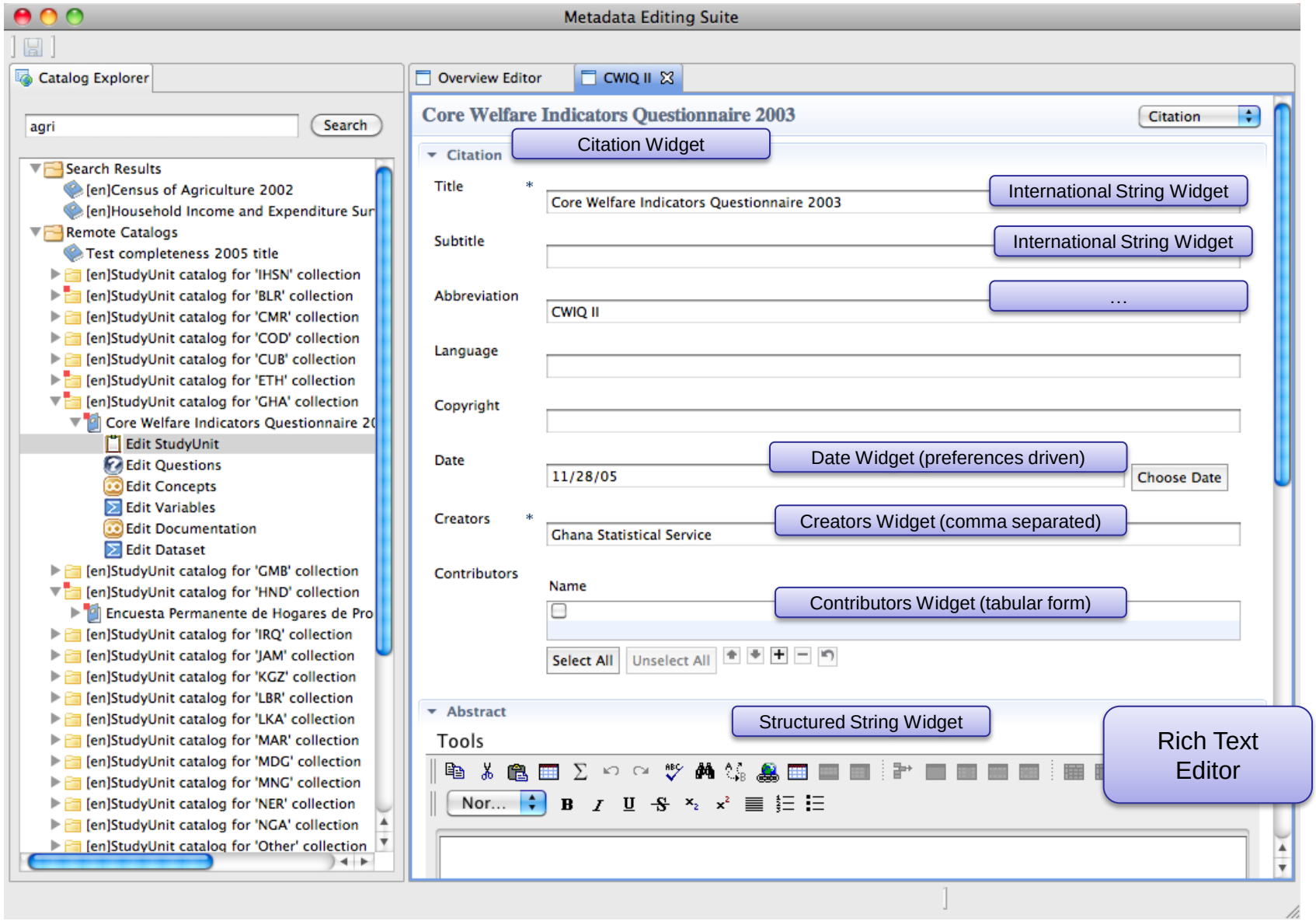
Some Design Techniques

- 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”

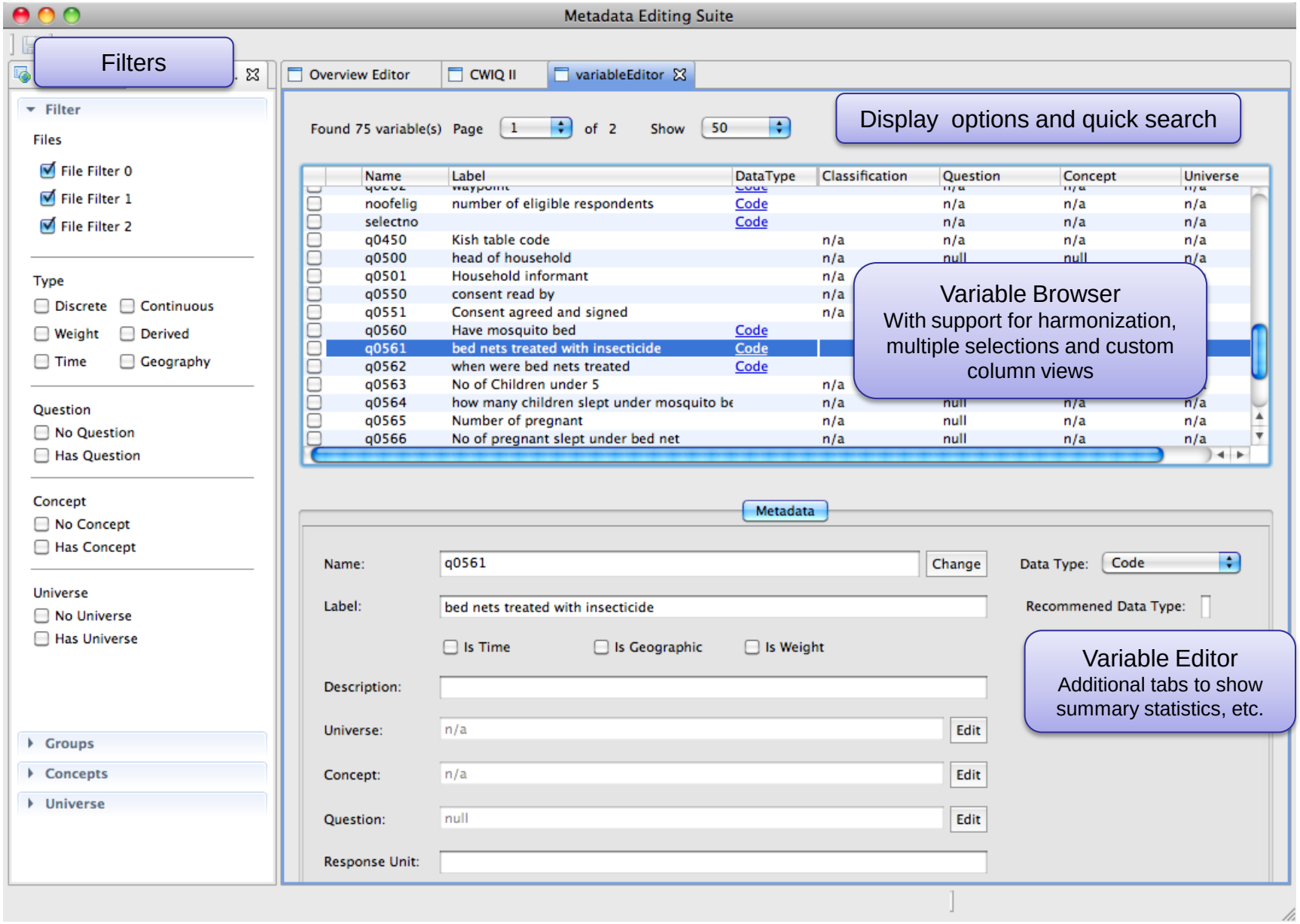
Catalog View



Study Editor



Variable Browser / Editor



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?

algenta
1428 Washington Ave S Ste 203
Minneapolis, MN 55454
USA
http://www.algenta.com

BRECKENHILL
305 Breckenridge Cres.
Ottawa, Ontario, K2W 1J3
Canada
http://wwww.breckenhill.com
contactus@breckenhill.com

ideas 2 evidence
Lyngveien 15
N-5101 Eidsvågneset
Norway
http://www.ideas2evidence.com

metadata technology
200 Prosperity Dr
Knoxville, TN 37923
USA
http://www.metadatatechnology.com
mtna@metadatatechnology.com