Supporting the sharing of longitudinal population health sciences data

Veerle Van den Eynden, MRC-DSS project

Medical Research Council, UK & UK Data Archive

IAASSIST, 7 June 2012
MRC data sharing vision

- New, quality research from existing data
- Visibility of data of individual studies and across studies
- Supporting studies to share research data
- Alignment and complementarity e.g. Cohort Resources Facility

MRC data policy
The MRC expects valuable data arising from MRC-funded research to be made available to the scientific community in a timely and responsible manner, with as few restrictions as possible so as to maximize the value of the data for research and for eventual patient and public benefit.
MRC Data Support Service project

- Guidance
  - Sharing of and access to research data, for population and patient studies
  - Data management planning
- Developing a sustainable MRC Research Data Gateway and Directory of Population and Patient Research Data
  - Discovery of research data, metadata, documentation, access
  - Engaging 15 largest MRC-funded (& co-funded) cohort studies
- Developing a data management toolkit to support effective data management and sharing
- Support MRC data managers network
  - Workshops
  - MRC-DATA-MANAGERS@JISCMAIL.AC.UK
- Alignment and partnership for standards, capability, infrastructure and funding
Data access and management Policies

1. Access policies are transparent
2. Independent oversight
3. PI roles are defined – & at arms length
4. Data Management policies & plans are transparent

Current DSS Phase 3

Clarifying MRC data policies
- New MRC guidance
- Harmonised with other funders

Support for PIs (& peer reviewers)
- Better Data Management Plans
- Better recognition for sharing and sharers

Better discovery of PHS studies
- Gateway prototype

Web-based discovery Gateway
1. MRC gateway to metadata directory
2. Extend from 4 to ~12 cohorts
3. Federated, with common standards
4. Integrate data discovery, access, curation and sharing workflows
MRC Research Data Gateway

- Discovery of MRC-funded studies, their data and variables
- Within and across studies (comparative discovery)
- Ambition - integrate data discovery, access, curation & sharing workflows
- Relationship to cohort studies:
  - Federated
  - Studies hold and govern access to data
  - Metadata developed by studies & published to Gateway directory
  - [Gateway facilitates & supports data access request]
- Alignment:
  - With CRF – facilitating data harmonisation and linkage
  - Internationally – e.g. P³G

www.datagateway.mrc.ac.uk
Gateway structure

- Prototype developed by consortium of Science and Technology Facilities Council, University College London and Oxford University, in collaboration with cohort case studies:
  - Avon Longitudinal Study of Parents and Children
  - National Survey for Health and Development
  - Southampton Women's Study
  - Whitehall II
- Drupal content management system
- Apache solr – indexing and searching
Metadata structure - nodes

- Study: a programme of research whereby data from and about individuals are collected and analysed
- Time period: a wave, sweep, time period or time point within a longitudinal study
- Data collection event: a survey, screening, interview series or clinic - the event through which research data were collected (one or many / time period)
- Variable: each data variable belongs to a particular study, phase and data collection event
- Resource: an information item for a study, time period, data collection event, e.g. questionnaire form, report, interview guide
  + subject categories
Gateway functionality

- Search across studies, time periods, data collection events, variables
- Filter and browse results – facets are ‘types’ of nodes and categories
- Explore groups of variables – study, time period, data collection event, category, previous/next variable
- Explore information (metadata)
- Select variables into basket
- Export basket for data access request to study
- Access controls to metadata
- Timeline of time periods across studies
- Forums
Welcome to the MRC Research Data Gateway

This is a prototype gateway for discovering MRC-funded studies and their datasets and variables.

At the heart of the gateway is the Population and Patient Research Data Directory, which holds detailed study and variables information.

All users of the gateway can:

- Discover cohort studies and their data collection events
- Find their procedures for data sharing

Registered users can also:

- Search and browse for variables of cohort studies
- Select and export variables to support a data sharing request

The directory does not contain research data themselves, only metadata and information on how to request access to data. Research data are held by the respective studies. Access to data can be requested from a study following its data sharing procedures.

The directory currently holds records for 33 studies and variables, time periods and data collection events for four studies:

- Avon Longitudinal Study of Parents and Children (39,358 variables for all 54 time points)
- National Survey for Health and Development (1137 variables for 2 out of 31 sweeps)
- Southampton Women’s Study (688 variables for the baseline and follow-ups during pregnancy and at birth)
- Whitehall II (4544 variables for the first 8 of the study’s 10 phases)

Posted January 27th, 2011 by administrator | Last updated January 18th, 2012
Search

Enter your keywords:

BMI

Search results

chek BMI                Error check status for BMI
  Error check status for BMI 115 months ... Number of responses: 7014
  Variable - alastairduncan - 2011-03-31 14:16 - 0 comments - 0 attachments - Variable

BMI                Body mass index
  Body mass index Phase 1 Screening Phase 1 Whitehall II ...
  Variable - alastairduncan - 2011-03-24 16:14 - 0 comments - 0 attachments - Variable

OBESITY              Degree of overweight - fromBMI
  Degree of overweight - fromBMI Phase 1 ...
  Variable - alastairduncan - 2011-03-24 16:14 - 0 comments - 0 attachments - Variable

XOBESITY              Degree of overweight - fromBMI
  Degree of overweight - fromBMI Phase 3 ...
  Variable - alastairduncan - 2011-03-24 16:14 - 0 comments - 0 attachments - Variable

dw042                BMI or Quetelet index
  BMI or Quetelet index 12 weeks gestation ... Number of responses: 12448
  Variable - alastairduncan - 2011-03-31 10:59 - 0 comments - 0 attachments - Variable

cf060                BMI at 4 mth
  BMI at 4 mth Children In Focus ... Number of responses: 1432
  Variable - alastairduncan - 2011-04-04 09:10 - 0 comments - 0 attachments - Variable

Add to My Variables
MRC Gateway: Standards

- Agreed standard metadata model
  - Mandatory and optional elements (for studies to map to)
- Moving to DDI as standard for metadata exchange
  - DDI-L
  - Metadata model mapped to DDI-L (3.1)
- MRC-defined standards for metadata quality
  - Challenge: Standardised concepts, topics,…
- Metadata publishing tools – repeatable, sustainable
  - Colectica Designer
  - Other tools
Metadata model agreed by community
Data /Metadata flows & governance

Metadata directory

- **Metadata publishing**
- **Derived variables**
- **Data request**
- **Data access**
- **Export basket of variables**
- **Search / browse**

Population & patient studies

Data

Metadata

- Study1 - wave1 - collection event 1 – var1
  - var2
  - var3
  - ....
  - collection event 2
  - wave2
  - wave3

- Study2 - wave1 - collection event 1 – var1
  - var2
  - var3
  - ....
  - collection event 2
  - wave2
  - wave3

Users / collaborators

MRC | Medical Research Council

Metadata curation, standardisation, added value
<table>
<thead>
<tr>
<th>DD13.1</th>
<th>Applied to Gateway</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>describes a cohort study</td>
<td></td>
</tr>
<tr>
<td>StudyUnit</td>
<td>describes a sweep / timeperiod</td>
<td></td>
</tr>
<tr>
<td>DataCollection</td>
<td>module, describing methodology, sampling,... of a StudyUnit or Group</td>
<td></td>
</tr>
<tr>
<td>Concept</td>
<td>reason for study; can also describe re-use potential or baskets of variables for particular research; can run across studies and groups, e.g. for shared variable baskets</td>
<td>not suitable for subject categories or variable relationships</td>
</tr>
<tr>
<td>Universe</td>
<td>defined for QuestionConstruct (StudyUnit)</td>
<td></td>
</tr>
<tr>
<td>QuestionScheme</td>
<td>defined for a collection event</td>
<td>sits within DataCollection module</td>
</tr>
<tr>
<td>QuestionText</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instruments</td>
<td>questionnaires, protocols</td>
<td>sits within DataCollection module</td>
</tr>
<tr>
<td>CategoryScheme</td>
<td>defined for each variable (or for groups of variables)</td>
<td>in LogicalProduct or ResourcePackage</td>
</tr>
<tr>
<td>CodeScheme</td>
<td>defined for each variable (or for groups of variables)</td>
<td>in LogicalProduct or ResourcePackage</td>
</tr>
<tr>
<td>Variable Scheme (in LogicalProduct)</td>
<td>describes attributes for each variable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Label</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UniverseRef</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ConceptRef</td>
<td></td>
</tr>
<tr>
<td></td>
<td>QuestionRef</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Representation (code scheme)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VariableGroup</td>
<td>link to Subject Categories; define relationship between variables</td>
</tr>
</tbody>
</table>

MRC | Medical Research Council
Metadata quality standards

0. Variables not yet coded or not yet in electronic format

1. Basic metadata: variable ID/label, brief variable description (often cryptic), variable values, number of responses

2. Detailed description (e.g. full question text, with question number, and page number in questionnaire; detailed description and protocol of an assessment/measurement)

3. Questionnaire / protocol available (e.g. PDF on website) to indicate question context, question sequence 3a PDF 3b PDF tagged with XMP metadata (XMP=extensible metadata platform, using DC) or DataBook metadata 3c questionnaire flow, caging, standard instrument indicated (e.g. via subject category

4. Source(s): reference to a questionnaire, protocol, or to other variable(s) from which a variable is derived (for derivations, cleaning)

5. Direct link to searchable PDF file to enable linking questions with place in questionnaire file (cfr QuestionBank)

6. Comments/notes: quality information about variable, e.g. changes in protocols, info on usability of variable, uncertainties in data,….

7. Categories / index / keywords – study-specific

8. Categories / index / keywords – standardised/generic, e.g. thesaurus, to be set by MRC DSS

9. Cross-language questionnaires, mapping translations
Principles of the Gateway: **Roles/responsibilities**

**Individual Studies**
- Curate & preserve research data
- Share data – minimise restrictions
- Balance study and user needs
- Develop quality data & metadata
- Publish metadata to gateway directory

**The Gateway**
- Set common standards
- Tools & guidance for metadata publishing
- Govern & host gateway & directory
- Support and guidance for studies
- Define clear goal & path
What next

• Test DDI3.1 xml ingest to Gateway directory
• Ingest more studies
• Co-develop with Cohort Resources Facility - Uniform Search Platform
• Co-develop with UK Data Archive
• coordinate with other data-discovery/search platforms, e.g. P3G, UK Biobank “Showcase” catalogue, planned UK Data Service (integrating social and economic data access), and government department initiatives
• Further develop functionality
• Further develop standardisation / harmonisation
Harmonisation of data & linkage

- harmonised variables / concepts

Discoverability, access, interoperability & standards

Platforms for data discovery, access and sharing for collaborative research