closer
Cohort & Longitudinal Studies Enhancement Resources

Why metadata is AWESOME!

Jon Johnson

May 21 2015

MRC Medical Research Council

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Overview

- A short digression on metadata
- Questionnaire inputs and outputs
- Metadata standards
- Implementing processes in a complex environment
- CLOSER Search Platform Development
- The integration of metadata and data management
- New research possibilities

closer
Cohort & Longitudinal Studies Enhancement Resources

Discovery & Classification

	Dewey Decimal System
000	Computer science and information
100	Philosophy and psychology
200	Religion and mythology
300	Social sciences
400	Language
500	Science and math
600	Technology
700	Arts and recreation
800	Literature
900	History and geography

closer
Cohort & Longitudinal Studies Enhancement Resources

Structure, navigation and meaning

+ Tables of contents
+ Indexes
+ Glossaries
+ References
+ Citations
+ Keywords

closer
Cohort & Longitudinal Studies Enhancement Resources

Describing the complex

closer
Cohort & Longitudinal Studies Enhancement Resources

Metadata, semantics and ontology

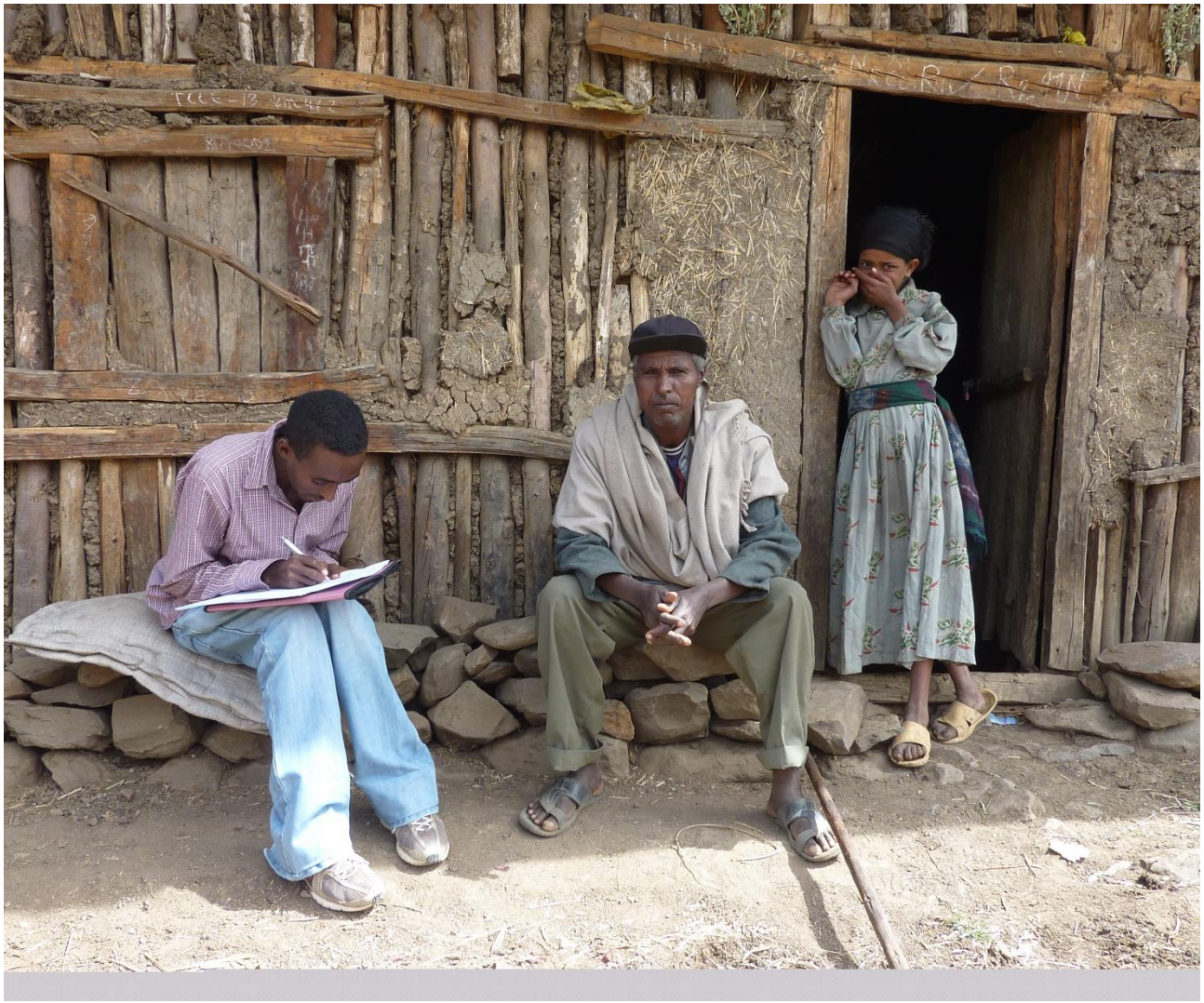
- Metadata is a mechanism for expressing the 'semantics' of information, as a means to facilitate information seeking, retrieval, understanding, and use.
- But **meaning** is as a 'locally constructed' artefact, [..], so that some form of agreement is required to maintain a common space of understanding.
- In consequence, metadata languages require **shared representations of knowledge** as the basic vocabulary from which metadata statements can be asserted.

Scilia, M. (2006) *Metadata, semantics, and ontology: providing meaning to information resources* Int. J. Metadata, Semantics and Ontologies, 1(1) p83

Metadata, semantics and ontology

- Ontology as considered in modern knowledge engineering is intended to convey that kind of **shared understanding**.
- In consequence, ontology along with (carefully designed) metadata languages can be considered as the foundation for a new landscape of information management.

Scilia, M. (2006) *Metadata, semantics, and ontology: providing meaning to information resources* Int. J. Metadata, Semantics and Ontologies, 1(1) p83



Lets start with the survey

closer
Cohort & Longitudinal Studies Enhancement Resources

What are survey questions trying to achieve

Accurate *Communication* & Accurate R*esponse*

Most important considerations are:

- *Language used*
- *Frame of reference*
- *Arrangement of questions*
- *Length of the questionnaire*
- *Form of the response*
  - *Dichotomous*
  - *Multiple choice*
  - *Check lists*
  - *Open Ended*
  - *Pictorial*

From Young, Pauline (1956) "Scientific Social Surveys & Research", 3rd Edition. Prentice Hall

closer
Cohort & Longitudinal Studies Enhancement Resources

What are we trying to capture

How the survey was *communicated* & how participants *responded*

Most important considerations are:

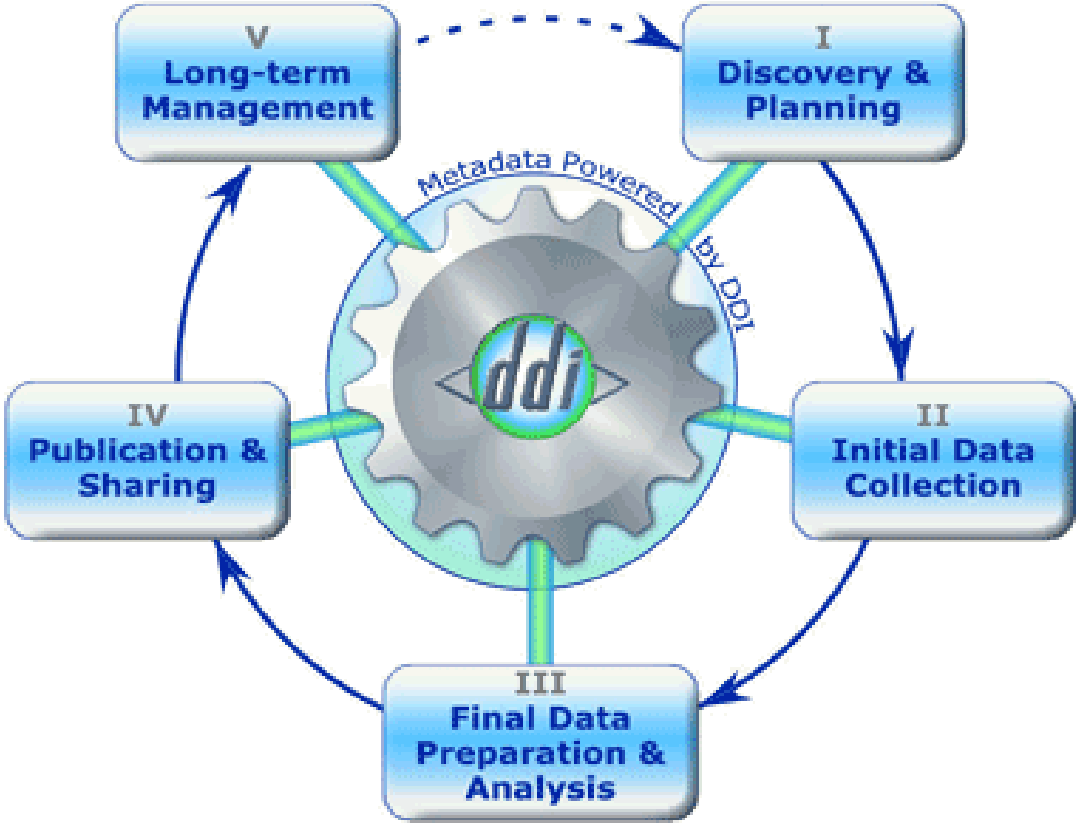
- *Language used in the questions*
- *Frame of reference*
- *Arrangement of questions*
- *Length of the questionnaire*
- *Form of the response*
  - *Dichotomous*
  - *Multiple choice*
  - *Check lists*
  - *Open Ended*
  - *Pictorial*
- *Who was asked*
- *Who responded*
- *Is the question asked related to another question*
- *Who was responsible for the collection*

closer
Cohort & Longitudinal Studies Enhancement Resources

A Common mechanism for communication

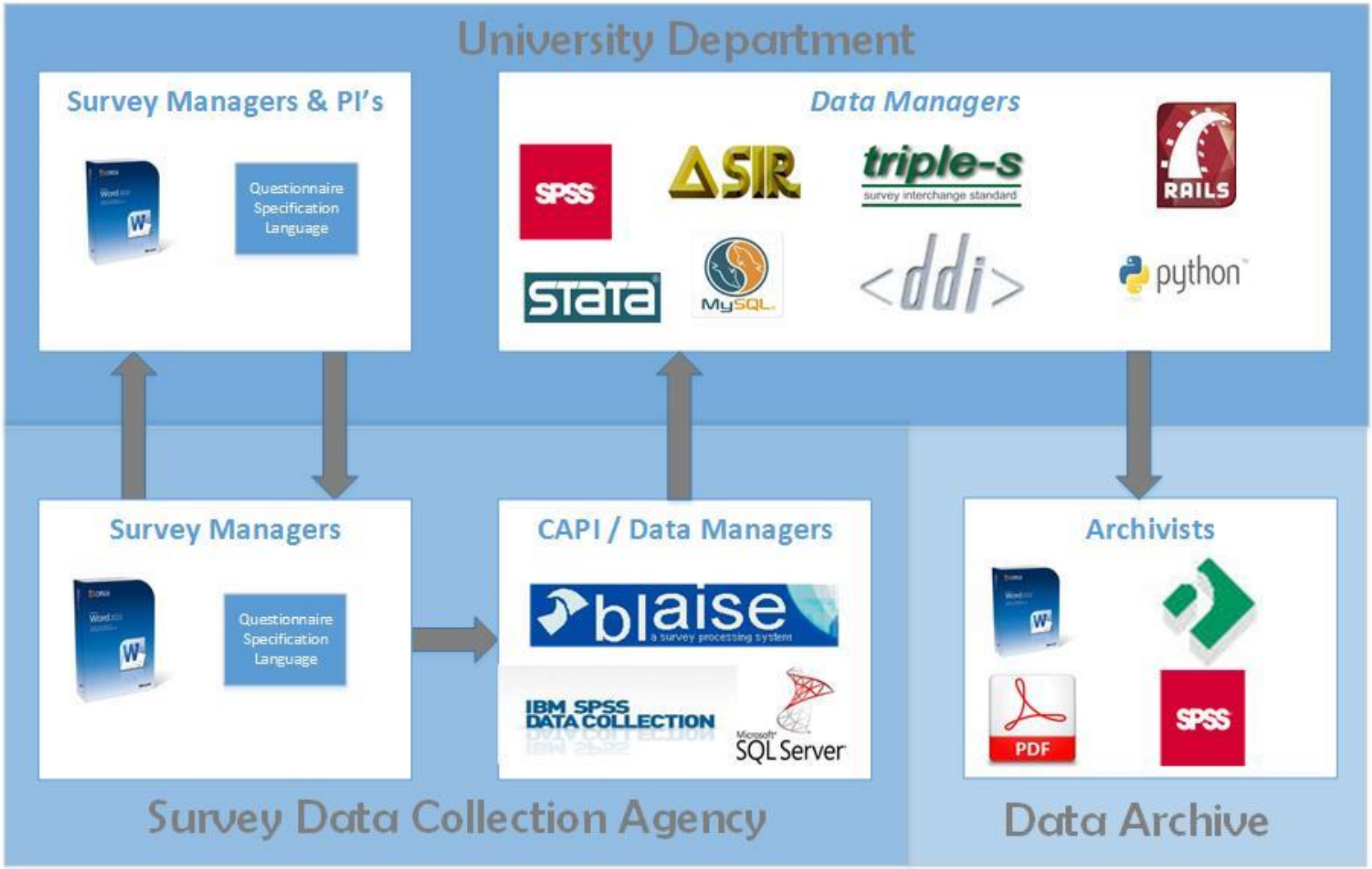
- Capture what was intended
  - What, where it came from and why
- Capture exactly what was used in the survey implementation
  - How, the logic employed and under what conditions
- To specify what the data output will be
  - That is mirrors what was captured and its source
- To keep the connection between the survey implementation through to the data received -> data management at CLS -> to the archive
- Generalised solution
  - So that is can be actioned efficiently and is self-describing
  - So that it can be rendered in different forms for different purposes

A Framework to work within



closer
Cohort & Longitudinal Studies Enhancement Resources

Current Longitudinal Survey Landscape



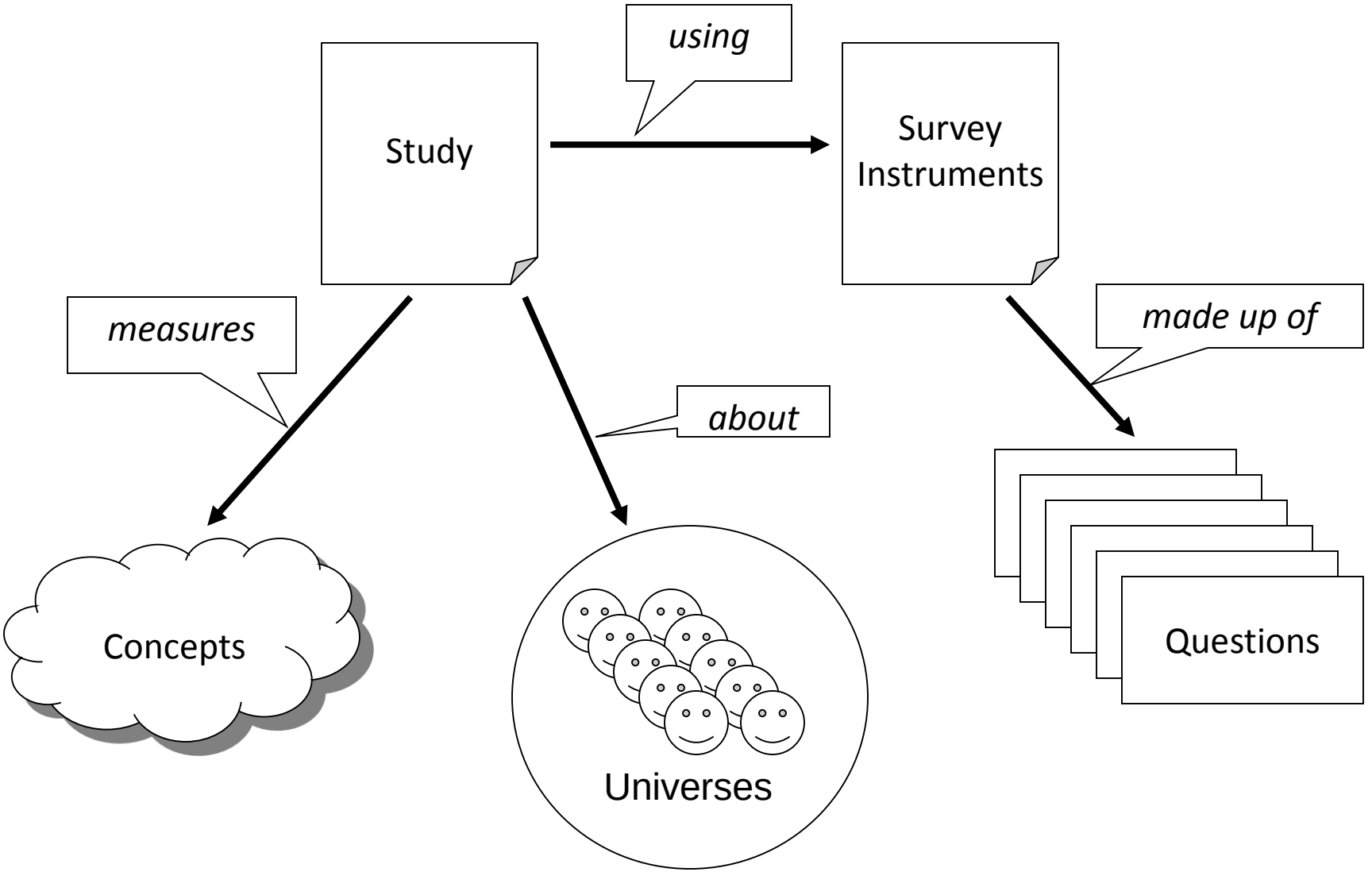
Barriers to sharing data and metadata

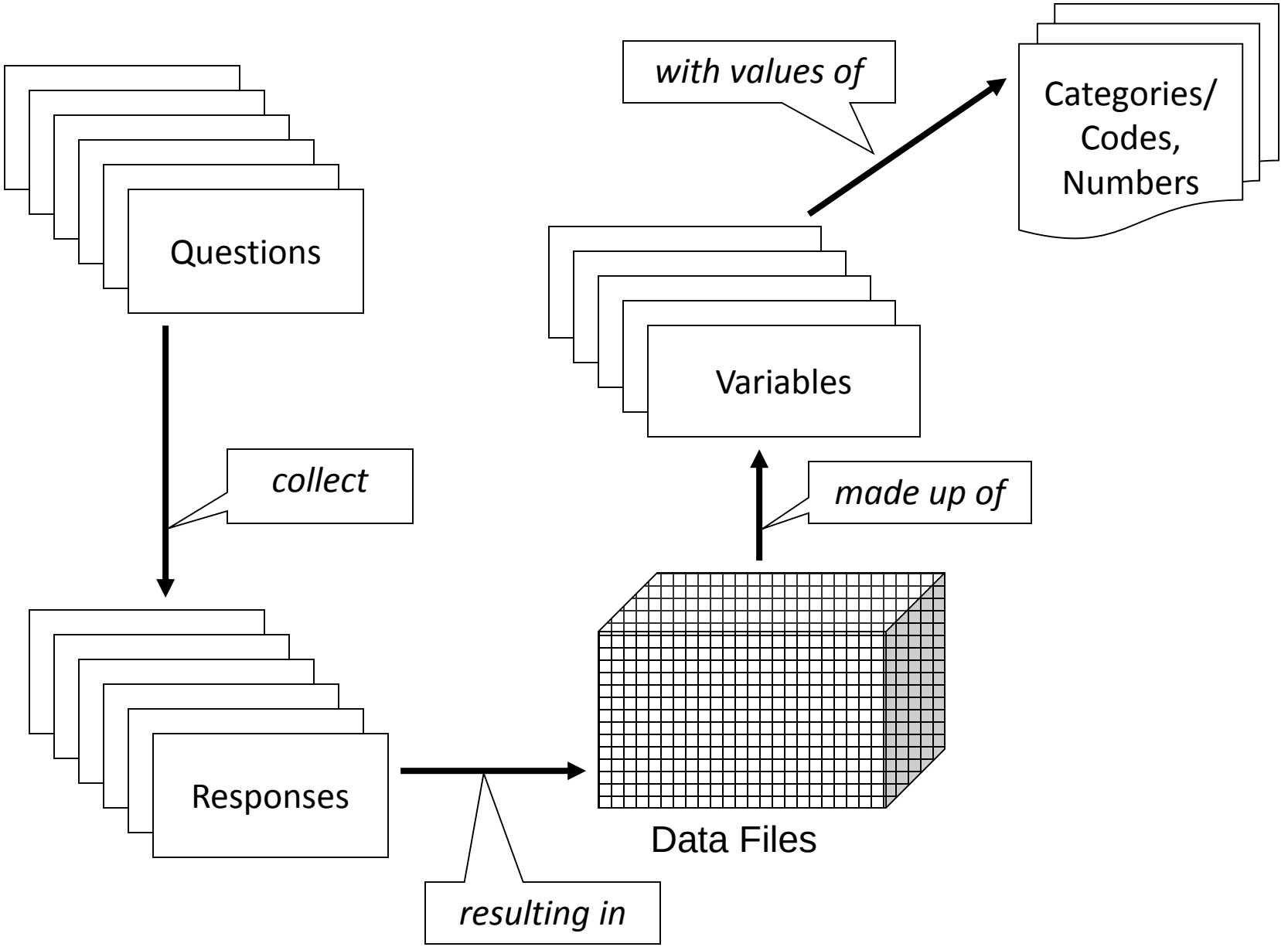
- Different agencies and clients have different systems
  - Taking over a survey from another agency often requires re-inputting everything
  - Questionnaire specification quality and format differences
  - Different clients have different requirements
- Barriers are also internal within organisations
  - Different disciplines have different attitudes to what is most important
  - Different departments speak different languages
  - Communication is always an issue
- Manual processes reduce transparency within and between organisations
- **Survey Metadata: Barriers and Opportunities" Meeting June 26, 2014, London sought to address some of these issues**

Adopting the standard

- The scale and complexity of the CAI instruments is a significant barrier to making the survey collection transparent and comprehensible to survey managers, researchers and analysts and for its subsequent data management
- CLS view the capturing of the implementation of the CAI .. in a standardised manner, to allow for version changes … during survey development and for later usage in data management and discovery as key output
- Survey contractors will be required to provide as a minimum a DDI-L XML compliant file of the CAI instruments within four months after the start of fieldwork …. and a mapping between survey questions and data outputs
- .. work with contractors to produce a 'human readable version' to improve usability of the questionnaire for end users

Learn DDI-L in 60 seconds

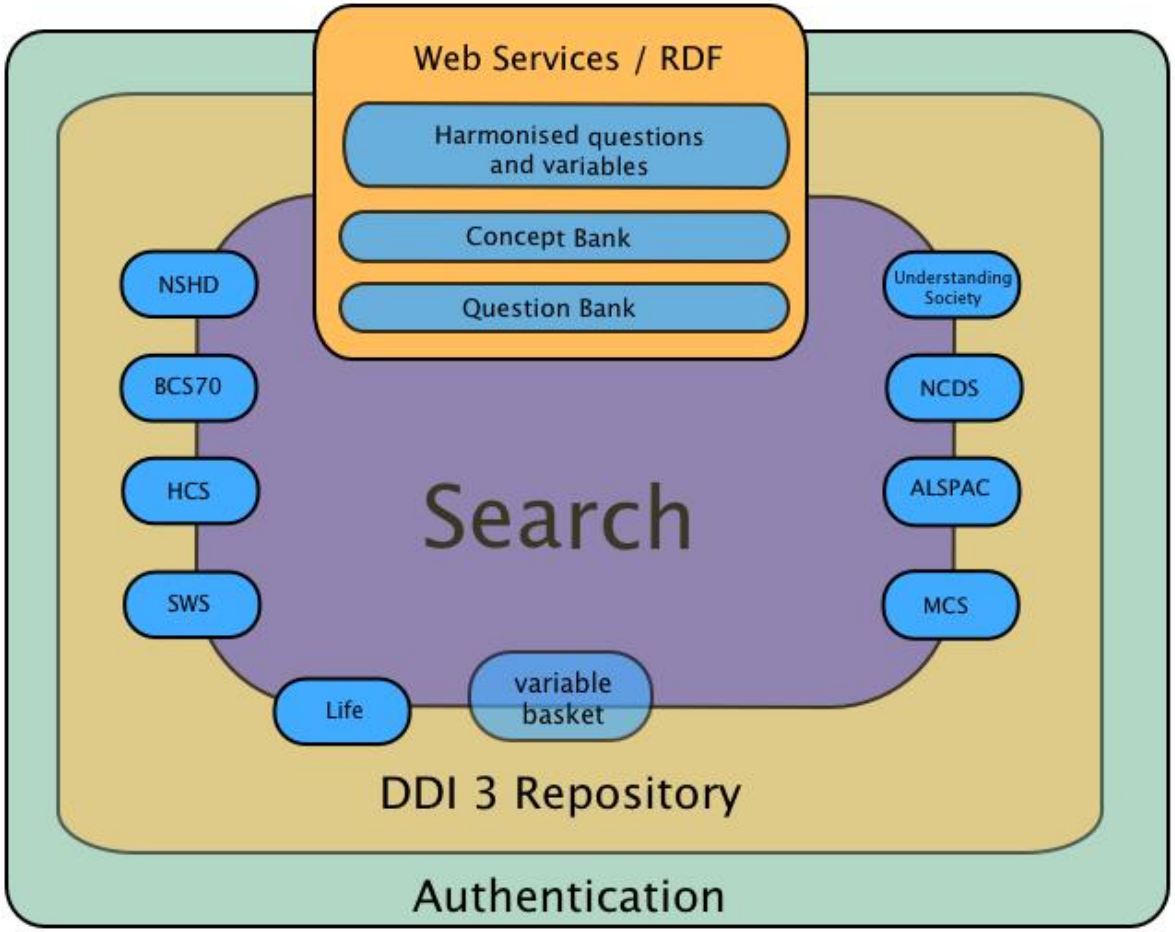




THAT'S PRETTY MUCH IT!

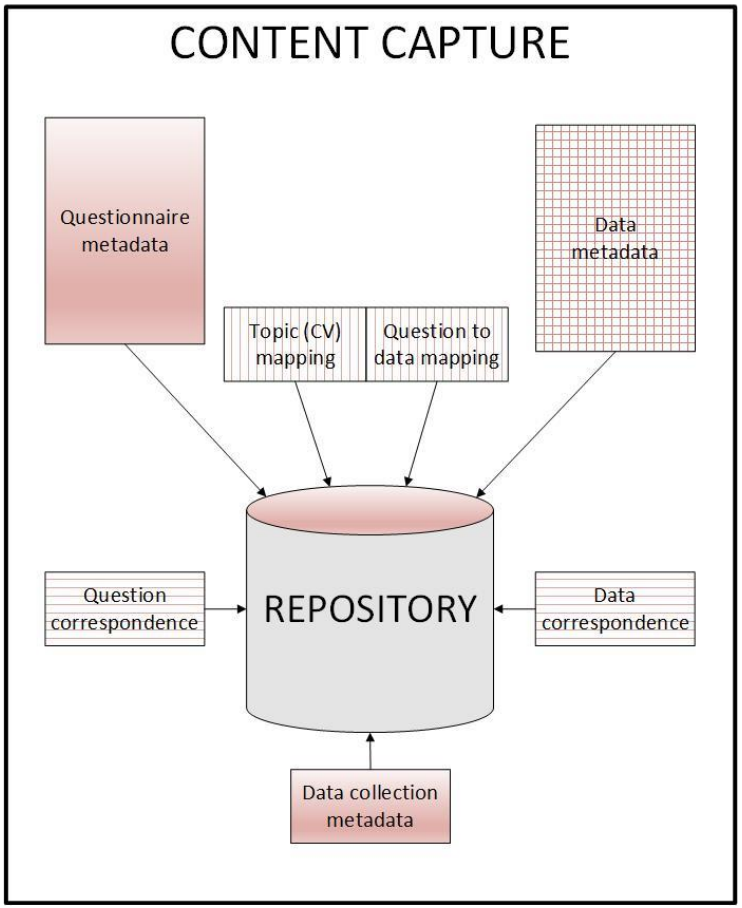
closer
Cohort & Longitudinal Studies
Enhancement Resources

CLOSER Metadata Search Platform



Building the Repository

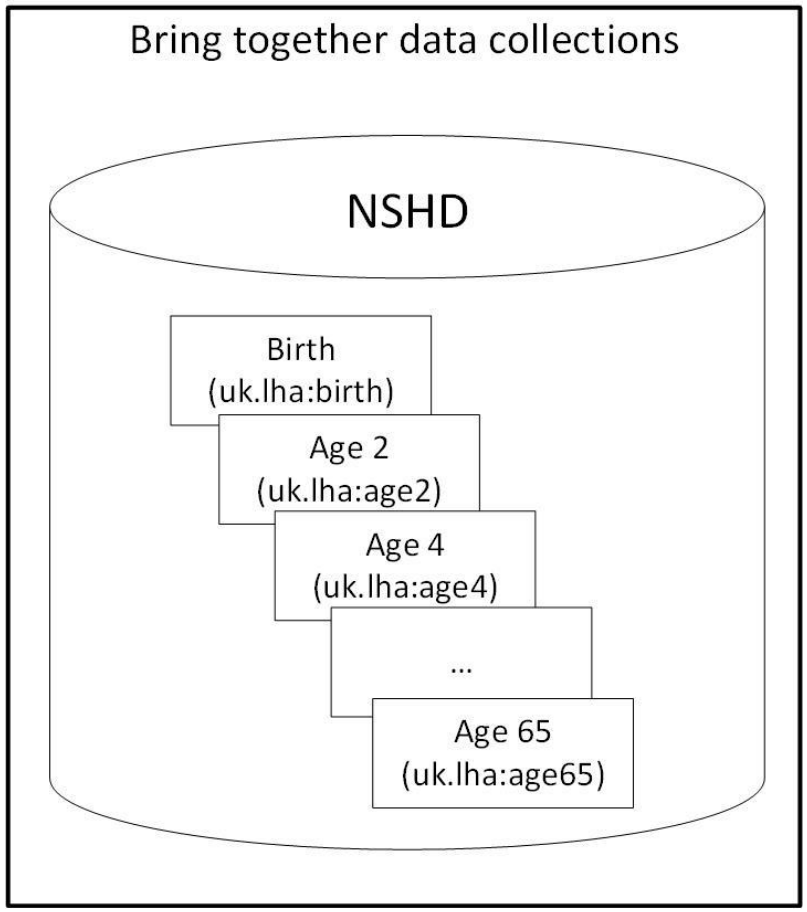
- Questionnaires from Studies
- Metadata extracted from data by studies
- Mapping by studies
- Correspondences by studies and CLOSER
- Reuse UKDA metadata and existing sources e.g. Life and Understanding Society
- Metadata Officer and Assistants input and co-ordination
- Ingest into Repository



closer
Cohort & Longitudinal Studies Enhancement Resources

Building the Repository

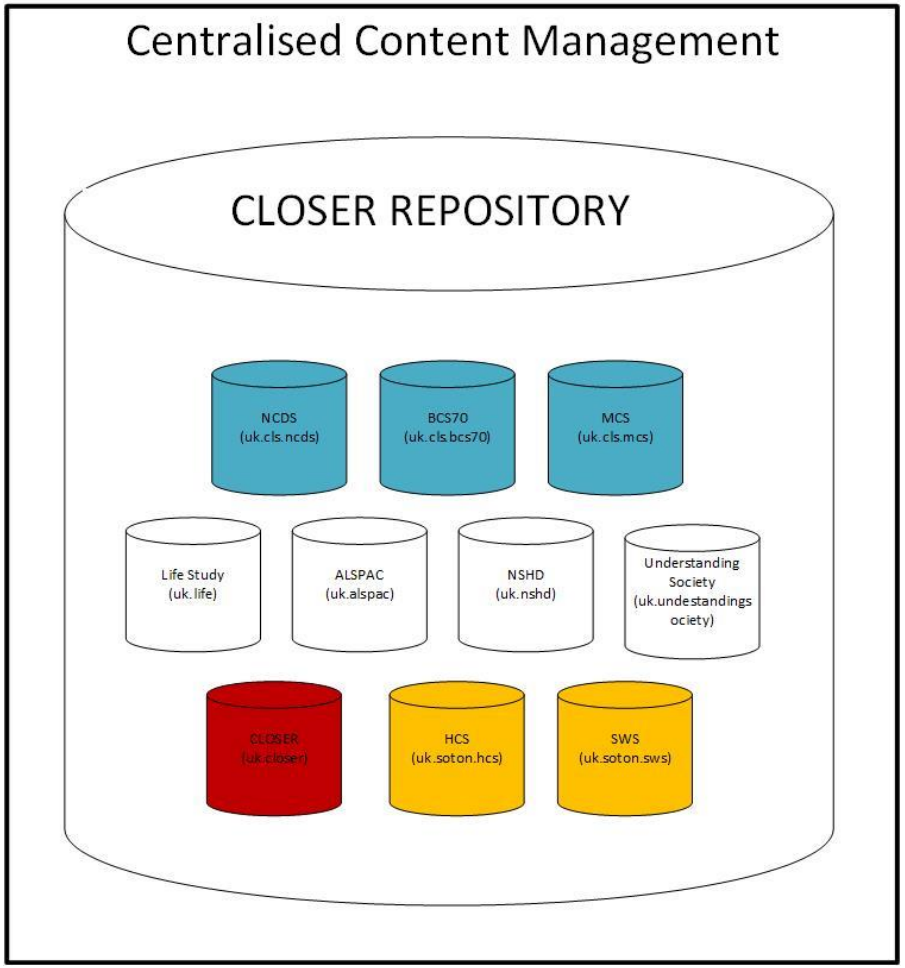
- Each data collection is treated as a separate entity
- Data collections are being added in sequentially (birth – latest)
- Each captured element, variable, question, instrument, study has its own **persistent identifier**
- Relationships are maintained by internal references
- Each study has its own identity



closer
Cohort & Longitudinal Studies Enhancement Resources

Building the Repository

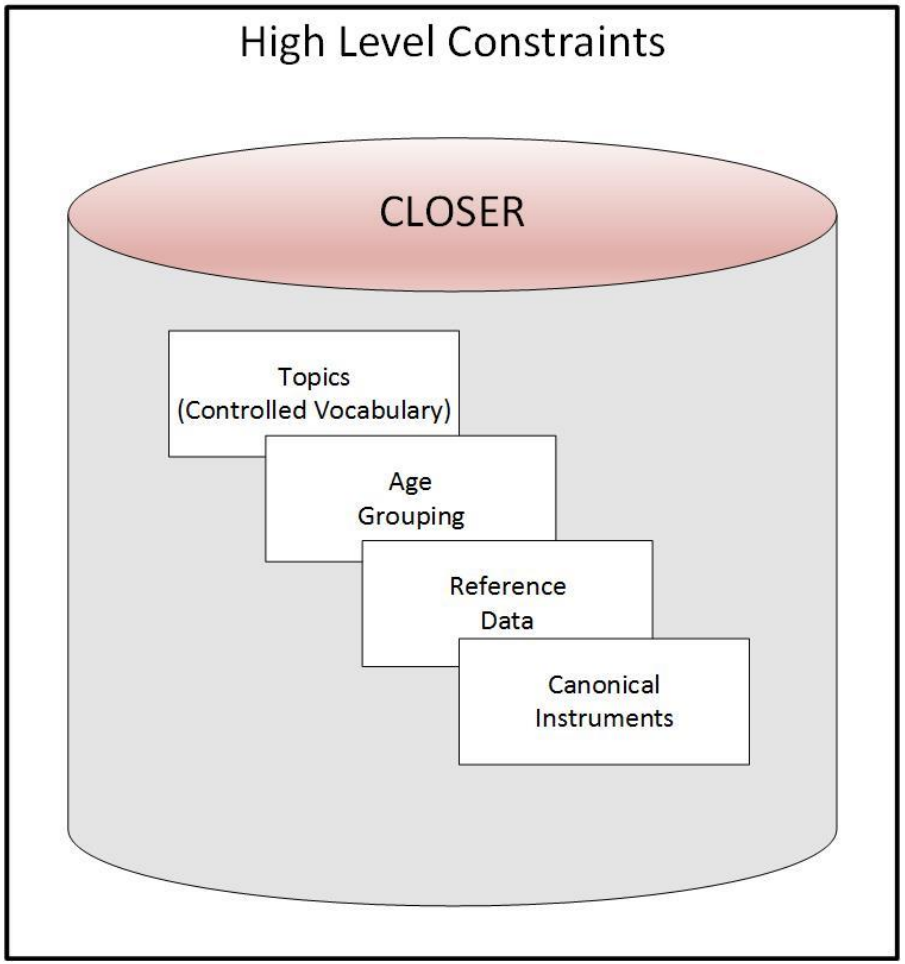
- Group studies by owner
- Connections between studies can be established to an item level
- Provenance is ‘built in’



closer
Cohort & Longitudinal Studies Enhancement Resources

Building the Repository

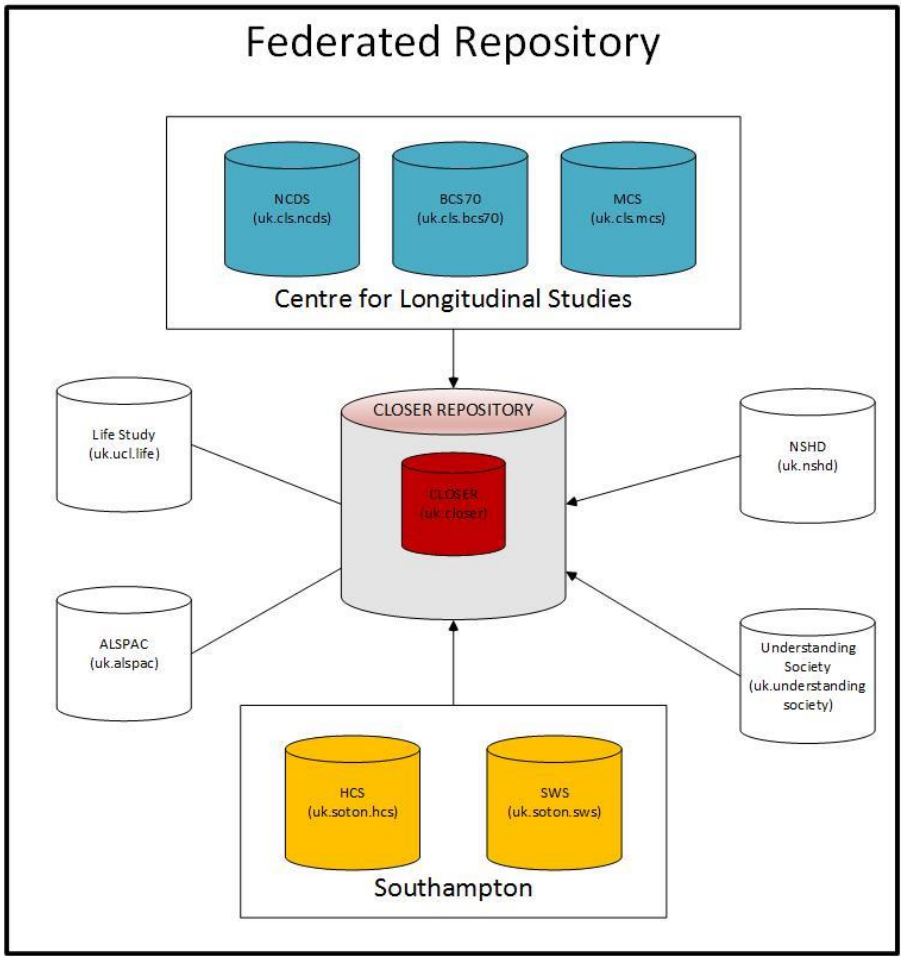
- Some things are maintained centrally
- Topics
  - Health, etc
- Life Stage(s)
- Reference data
  - Occupation coding
  - Geography
  - Schemes
- Canonical instruments
  - GHQ
  - SDQ
  - Rutter



closer
Cohort & Longitudinal Studies
Enhancement Resources

Building the Repository

- Ownership is returned to the studies
- Control by studies of what is pushed to the centre?
- Long term maintenance and management planning
  - Resourcing
  - Training
  - Capacity planning
- New Studies can be brought in



closer
Cohort & Longitudinal Studies Enhancement Resources

Metadata management -> Data Management

- All objects in a DDI have a URN.
- These are intended to serve as persistent, location-independent identifiers, allowing the simple mapping of namespaces into a single URN namespace.
- The existence of such a URI does not imply availability of the identified resource, but such URIs are required to remain globally unique and persistent, even when the resource ceases to exist or becomes unavailable
- urn:ddi:DDIAgencyID:BaseID:Version
- e.g. urn:ddi:uk.closer:thingamajig:1.0.0

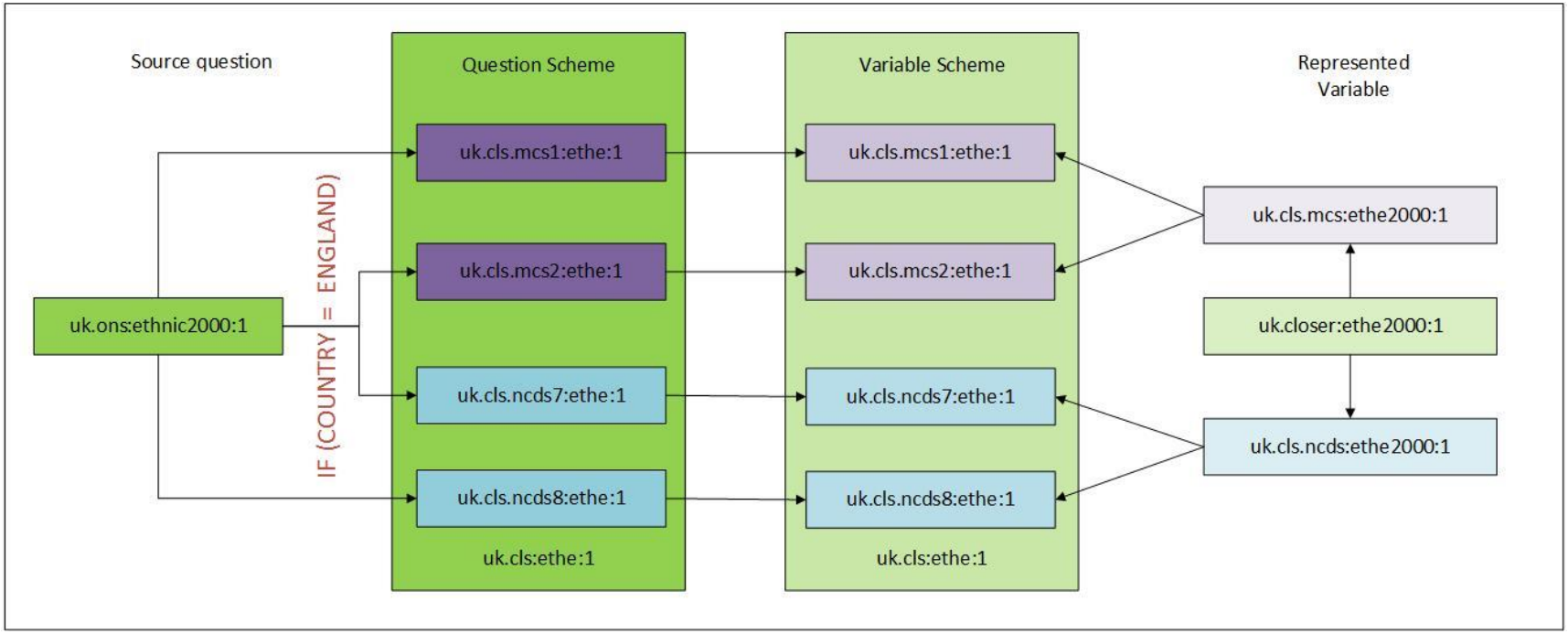
closer
Cohort & Longitudinal Studies Enhancement Resources

Information, Meaning and Relationships

- ETHNIC
  - White / Black / Asian / Other
- Universe
  - ETHNICE == respondents England
  - ETHNICN == respondents (N Ireland)
- Concept
  - "self defined ethnic identity"
- Based on
  - 2000 ONS self defined ethnic identity
- Equal to
  - 2010 ONS self defined ethnic identity
- Comparison
  - ETHNICE (3) == ETHNICN (2)
- Agency
  - uk.ons:ethnic2000:1.0 = ETHNIC 2000
  - uk.ons.ethnic2010.1.0 = ETHNIC 2010

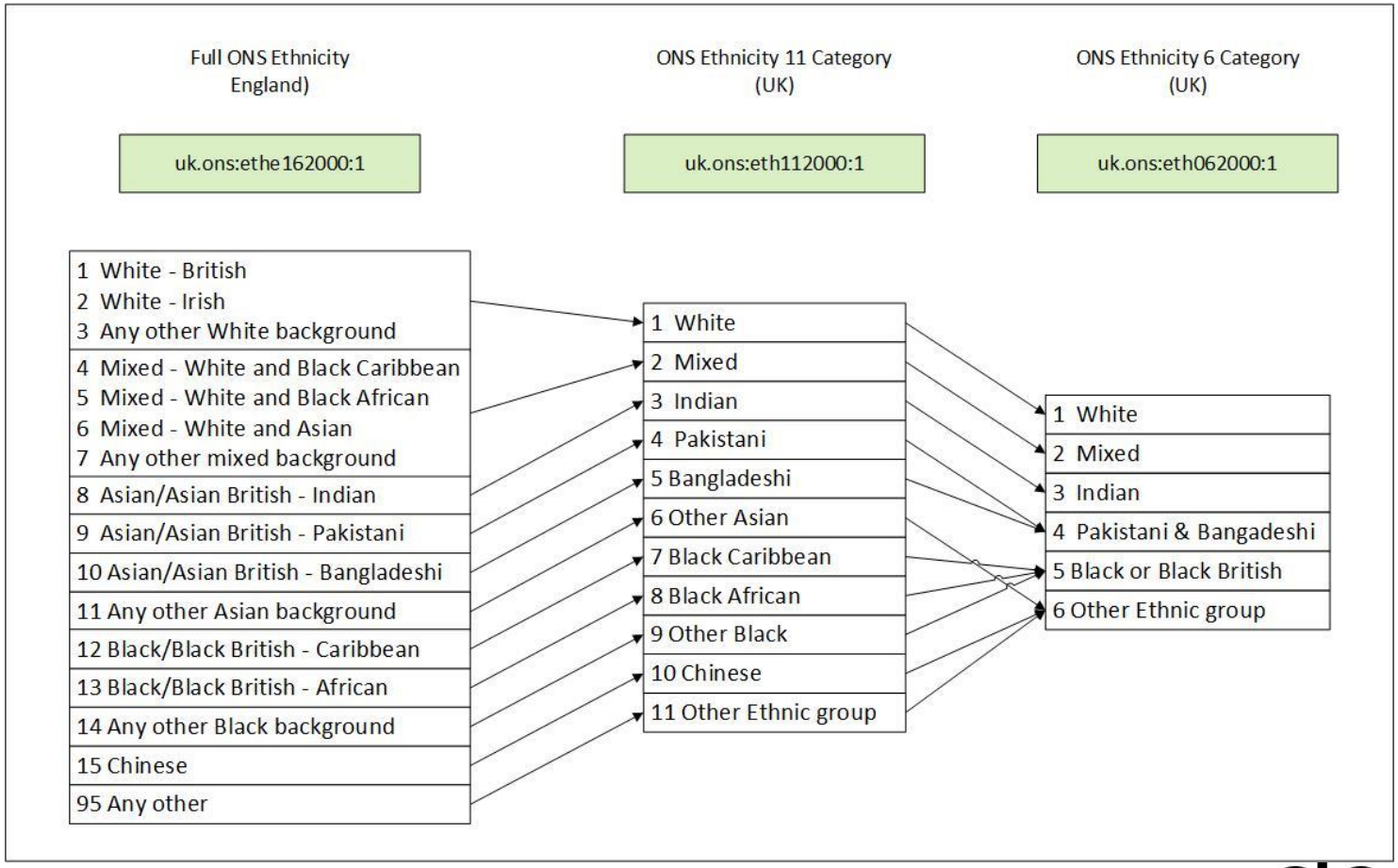
- **White**
- 1. English/Welsh/Scottish/Northern Irish/British
  2. Irish
  3. Gypsy or Irish Traveller
  4. Any other White background, please describe
- **Mixed/Multiple ethnic groups**
- 5. White and Black Caribbean
  6. White and Black African
  7. White and Asian
  8. Any other Mixed/Multiple ethnic background, please describe **Asian/Asian British**
- 9. Indian
  10. Pakistani
  11. Bangladeshi
  12. Chinese
  13. Any other Asian background, please describe **Black/ African/Caribbean/Black British**
- 14. African
  15. Caribbean
  16. Any other Black/African/Caribbean background, please describe **Other ethnic group**
- 17. Arab
  18. Any other ethnic group, please describe

Question and variable organisation



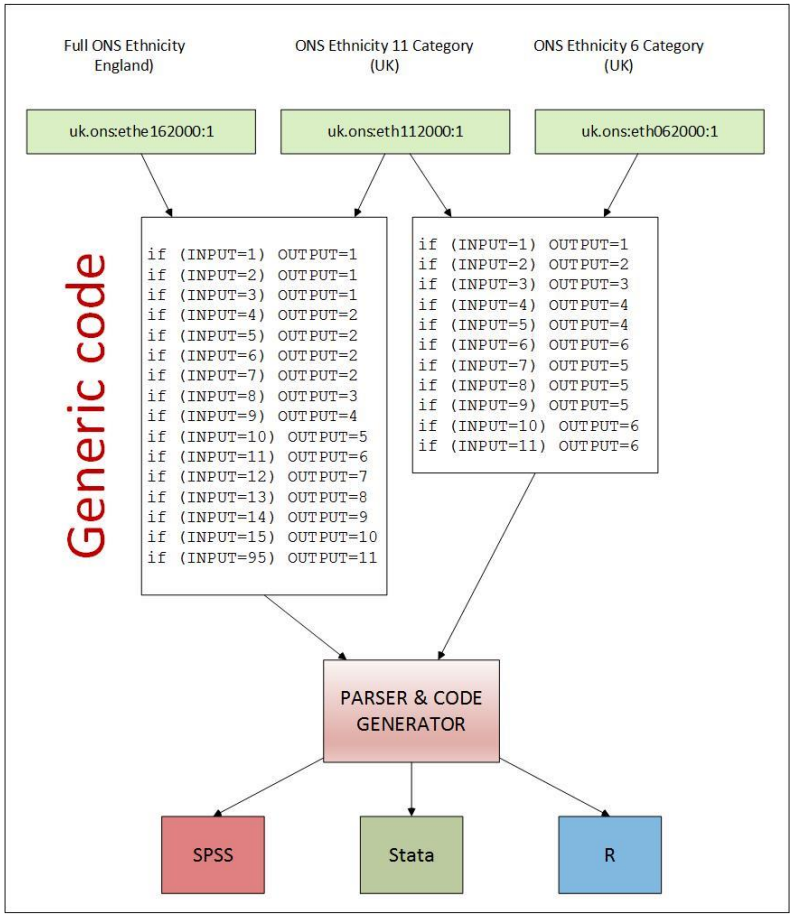
closer
Cohort & Longitudinal Studies Enhancement Resources

Code List Mapping



closer
Cohort & Longitudinal Studies Enhancement Resources

Metadata Code generation



Use Cases

- Harmonisation
- Common code base from same metadata
- Platform independence
- Reproducibility of outputs

closer
Cohort & Longitudinal Studies Enhancement Resources

Understanding change

ICD9 to ICD10

Is There a One-to-One Match Between ICD-9-CM and ICD-10?

No, there is not a one-to-one match between ICD-9-CM and ICD-10, for which there are a variety of reasons including:

➤ There are new concepts in ICD-10 that are not present in ICD-9-CM;

➤ For a small number of codes, there is no matching code in the GEMs;

➤ There may be multiple ICD-9-CM codes for a single ICD-10 code; and

➤ There may be multiple ICD-10 codes for a single ICD-9-CM code.

- Comparison mapping between different codes (“things that mean something”)
- Concepts e.g. laterality in ICD10
- Processing instructions leverage meaning and concepts

closer
Cohort & Longitudinal Studies Enhancement Resources

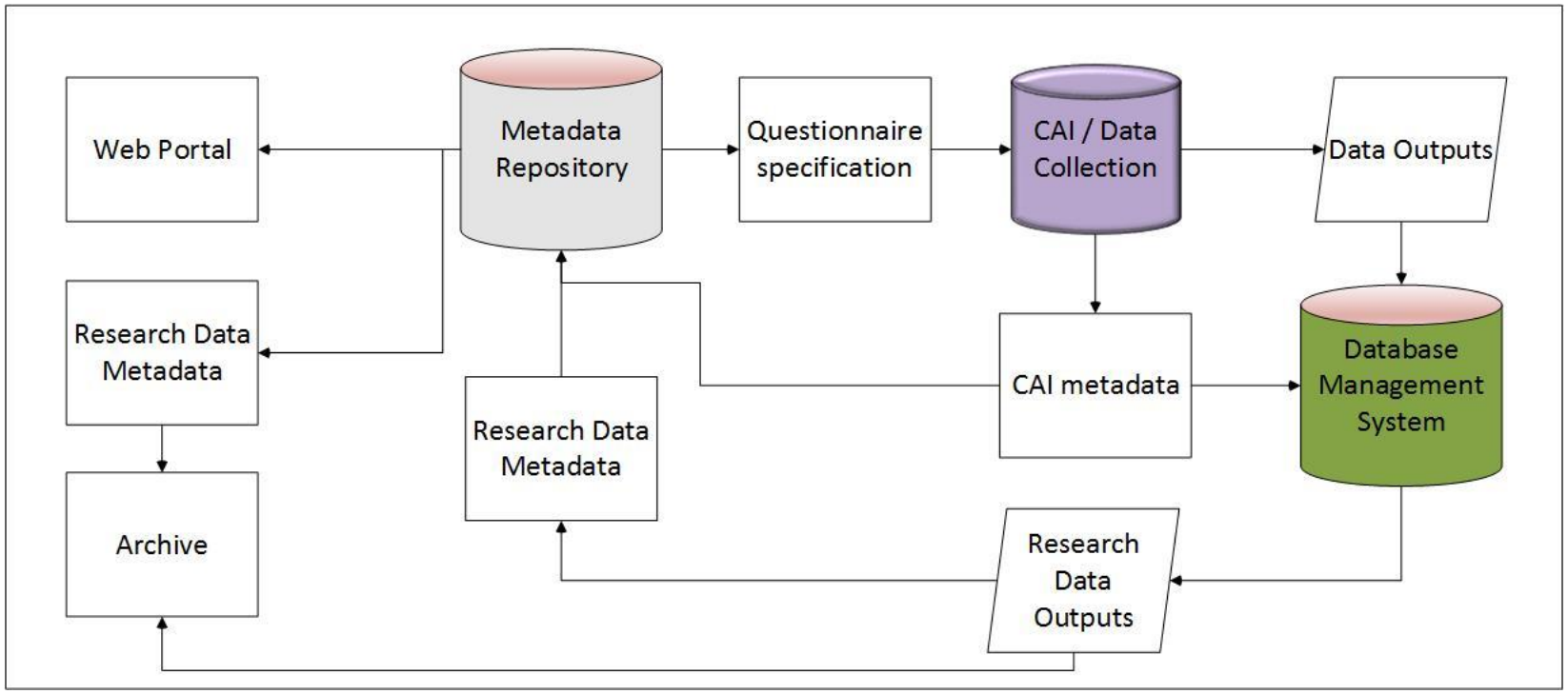
Tracking Version

ICD-10-CM to ICD-9-CM GEM entry for "Toxic effect of lead, cause undetermined"

2014 entry	Updated 2015 entry	Comment
T56.0X4A Toxic effect of lead and its compounds, undetermined, initial encounter To **Choice List 1** To **984.9** Toxic effect of unspecified lead compound **AND** **Choice List 2** To **980.9** Toxic effect of unspecified alcohol	**T56.0X4A** Toxic effect of lead and its compounds, undetermined, initial encounter To **Choice List 1** To **984.9** Toxic effect of unspecified lead compound **AND** **Choice List 2** To **E980.9** Poisoning by other and unspecified solid and liquid substances, undetermined whether accidentally or purposely inflicted	Typographical error. The E was missing from the external cause ICD-9-CM code in choice list 2.

closer
Cohort & Longitudinal Studies Enhancement Resources

Metadata Driven Pipeline



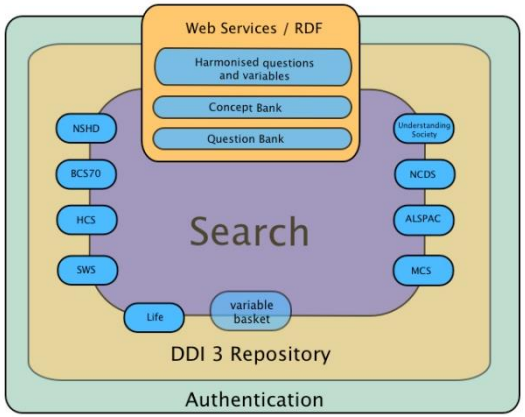
closer
Cohort & Longitudinal Studies Enhancement Resources

Let's DISCO

```
PREFIX disco:   <http://rdf-vocabulary.ddialliance.org/discovery#>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX skos:    <http://www.w3.org/2004/02/skos/core#>

SELECT COUNT(?universe) AS ?no ?universeDefinition
WHERE {
    ?universe a disco:Universe .
    ?universe skos:definition ?universeDefinition.

    FILTER(langMatches(lang(?universeDefinition ), "EN"))
    FILTER regex(?universeDefinition , "SEARCHWORD", "i")
}
GROUP BY ?universeDefinition
ORDER BY DESC(?no)
LIMIT 10
```



http://ddi-rdf.borsna.se/examples/gexf/

closer
Cohort & Longitudinal Studies Enhancement Resources

Some final thoughts

- Reduction in manual processes
- Enables distributed data collection
- Enables distributed research
- Increased quality of documentation of data collection
  - Raises visibility of needs
  - Encourages users to better understand
    - the data and
    - the data collection process
- New tools to think in more interesting ways can be built

closer
Cohort & Longitudinal Studies Enhancement Resources