

Design and generation of Linked Clinical Data Cube

Laurent Lefort and Hugo Leroux

1st Workshop on Semantic Statistics, 22 October 2013

CSIRO COMPUTATIONAL INFORMATICS

www.csiro.au



Australian Imaging Biomarkers and Lifestyle data

Problem – Solution – Remaining challenges

- Data collected for the AIBL study
 - (aibl.csiro.au)



- For the early detection of Alzheimer's Disease
- Protocol aligned with Alzheimer's Disease Neuroimaging Initiative (ADNI)
 - Plus nutrition, lifestyle
- Microdata collected via electronic data capture tool (OpenClinica)
- To be consumed by researchers
 - 1) discovery + data quality assessment (fix)
 - 2) production of publishable results
- Original format (export format): CDISC ODM
 - Clinical Data Interchange Standards Consortium
 - Operational Data Model

AIBL Protocol and e-data capture

		Screening	Baseline	18 Months	36 Months	54 Months	72 Months	...
Vital Signs		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Blood		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Neuropsychological t.		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Demographics	<input type="checkbox"/>	<input type="checkbox"/>						
Medications		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
PiB PET scan and MRI for ¼		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Diagnostic Summary	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Diet and lifestyle Q.		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Vital Signs V1.0

Event: Collection 1 (20090913) Site: 11
 Study: AIBL Study Age At Enrollment: 74 Years (21 Days)
 Date of Birth: 10/01/1947
 Site: 11A Person ID: 1
 Recruitment Status: 100 Review Date: 20090913

Recency: None Updated Revision Proposed Closed Not Applicable

There are issue(s) with your submission. The data has NOT been saved. See below for details.

Issue(s) Found: Invalid, duplicate, or empty values submitted. [Update status to a numeric value between 01 and 100]

WV: 0001 Select to jump

Title: Vital Signs

Page: 1 Mark CAP Complete

Height: 155 cm (in)

Weight: 65 kg (lb)

Systolic Blood Pressure: 125 mm (mm Hg)

Diastolic Blood Pressure: 80 mm (mm Hg)

Heart Rate: 51 bpm

Abnormal: 0 (yes)

Vital signs comments:

Body Mass Index: 27.33

 Mark CAP Complete

Medical History V1.2 0006

Click the box next to an input to enter/overwrite values. Please note that you can only save the value if CAP data entry has already started.

Medical History

Medical History

Medical History

Medical History Details

Is the participant currently taking any medication?

Medication	Dose	Frequency	Length of Time Taken	For what condition are you taking this medication?
Lorazepam	1mg	2 times per day	4 years	
Plavix	75mg	1 time per day	4 years	
Lisinopril	10mg	1 time per day	7 years	
Lasix	40mg	1 time per day	4 years	
Kabool	1mg	1 time per day	4 years	
Aspirin	100mg	1 time per day	15 years	

Has the participant had any changes to their medication since baseline?

Examples of forms (OpenClinica)

Vital Signs V1.0

CRF Header Info

Event:	Collection 1 (22/07/2011)	Sex:	M
Study:	AIBL Study	Age At Enrollment:	24 Years - 21 Days
Site:	N/A	Date of Birth:	01/07/1987
Interviewer Name: *	root	Person ID:	1
		Interview Date: *	22/07/2011

Discrepancy Notes on this CRF:

New	Updated	Resolution Proposed	Closed	Not Applicable
0	0	0	0	0

There are issue(s) with your submission. The data has NOT been saved. See below for details.

- **[Blood Pressure Systolic expects a numeric value between 70 and 199]**
- **[Weight expects a numeric value between 40 and 200]**

VitalSi...(0/8) -- Select to Jump --

Title: Vital Signs

Page: 1 Mark CRF Complete **Save** **Exit**

Height: 150 (cms)

Weight: 39 (kg)

Systolic Blood Pressure: 69 (mm of Hg)

Diastolic Blood Pressure: 40 (mm of Hg)

Heart Rate: 51 (bpm)

Abdominal Circumference: 40 (cms)

Vital Signs Comments

Body Mass Index: 17.33

Return to top Mark CRF Complete **Save** **Exit**

Medical History V1.2

CRF Header Info

Click the flag icon next to an input to enter/view discrepancy notes. Please note that you can only save the notes if CRF d

Exit

Medical...(0/131) **Medicat...(0/25)** -- Select to Jump --

Title: Medications

Subtitle: Medications

Page: 2

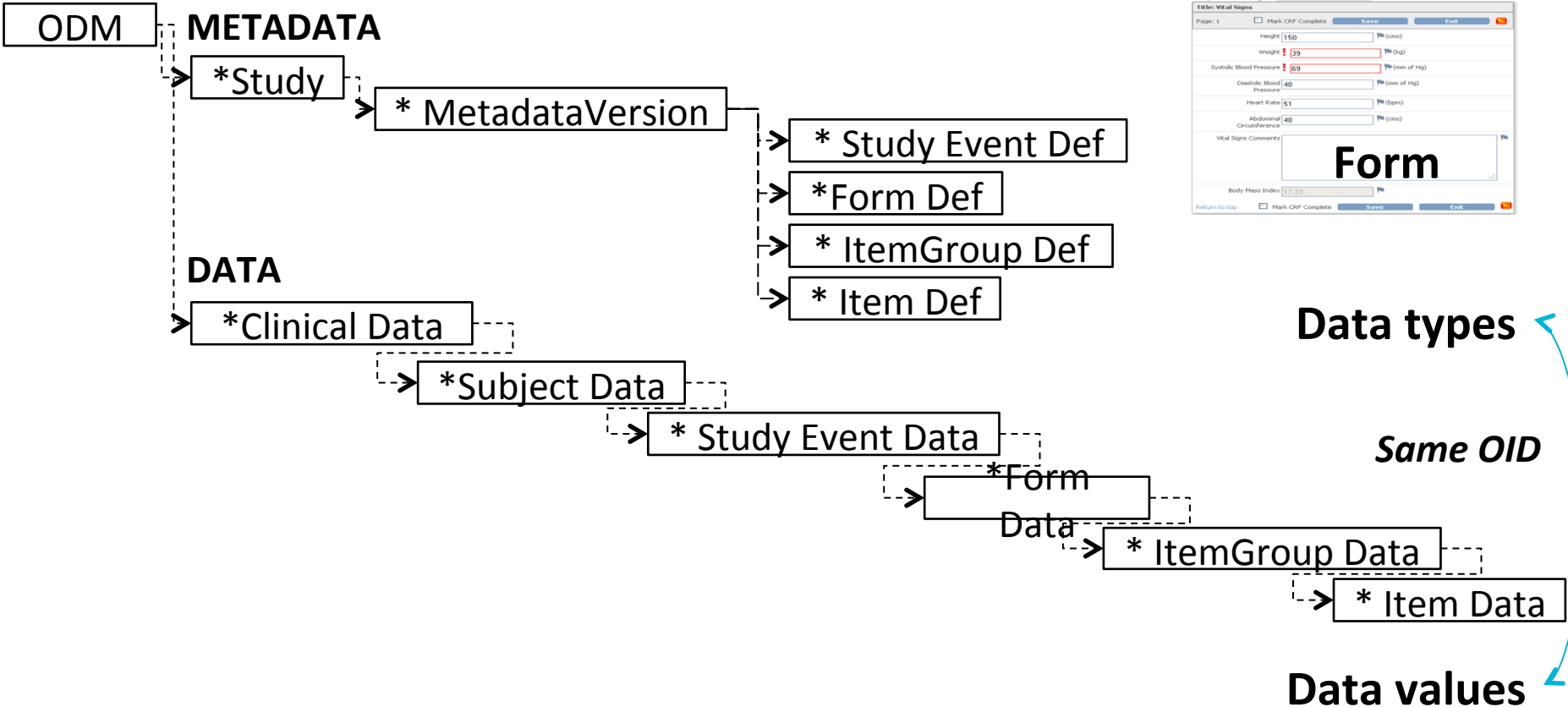
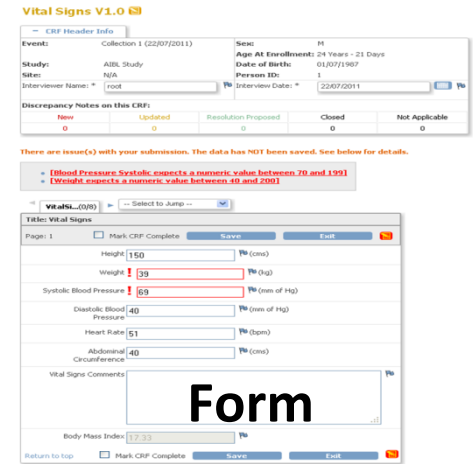
Medications Details

Is the participant currently taking any medication? Yes

Medications

Name:	Dose:	Frequency:	Length of Time Taken:	For what condition are you taking this medication?
Lenoxin	125 (mg)	2 (day)	4 years	
Plavix	10 (mg)	1 (day)	4 years	
Lipitor	(mg)	1 (day)	7 years	

CDISC ODM XML Schema

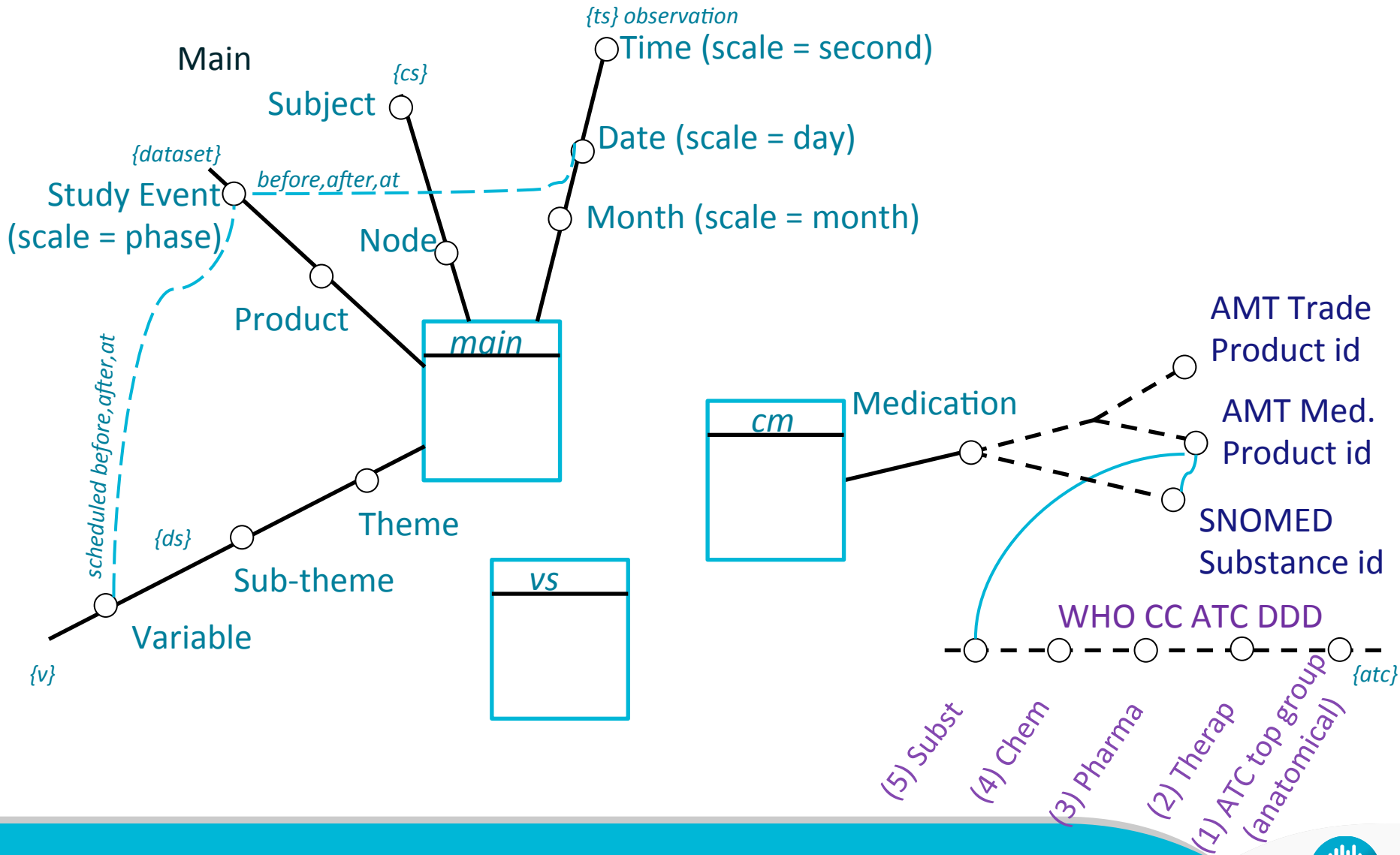


Current use of AIBL data

Problem – Solution – Remaining challenges

- Conversion in tabular format (Excel or CSV)
- Browser-based exploration tool
- Additional processing via Excel or R or ...
- (different toolset for AIBL and ADNI)

Primary motivation: add new dimensions



Is this a Semantic Stats problem?

Problem – **Solution** – Remaining challenges

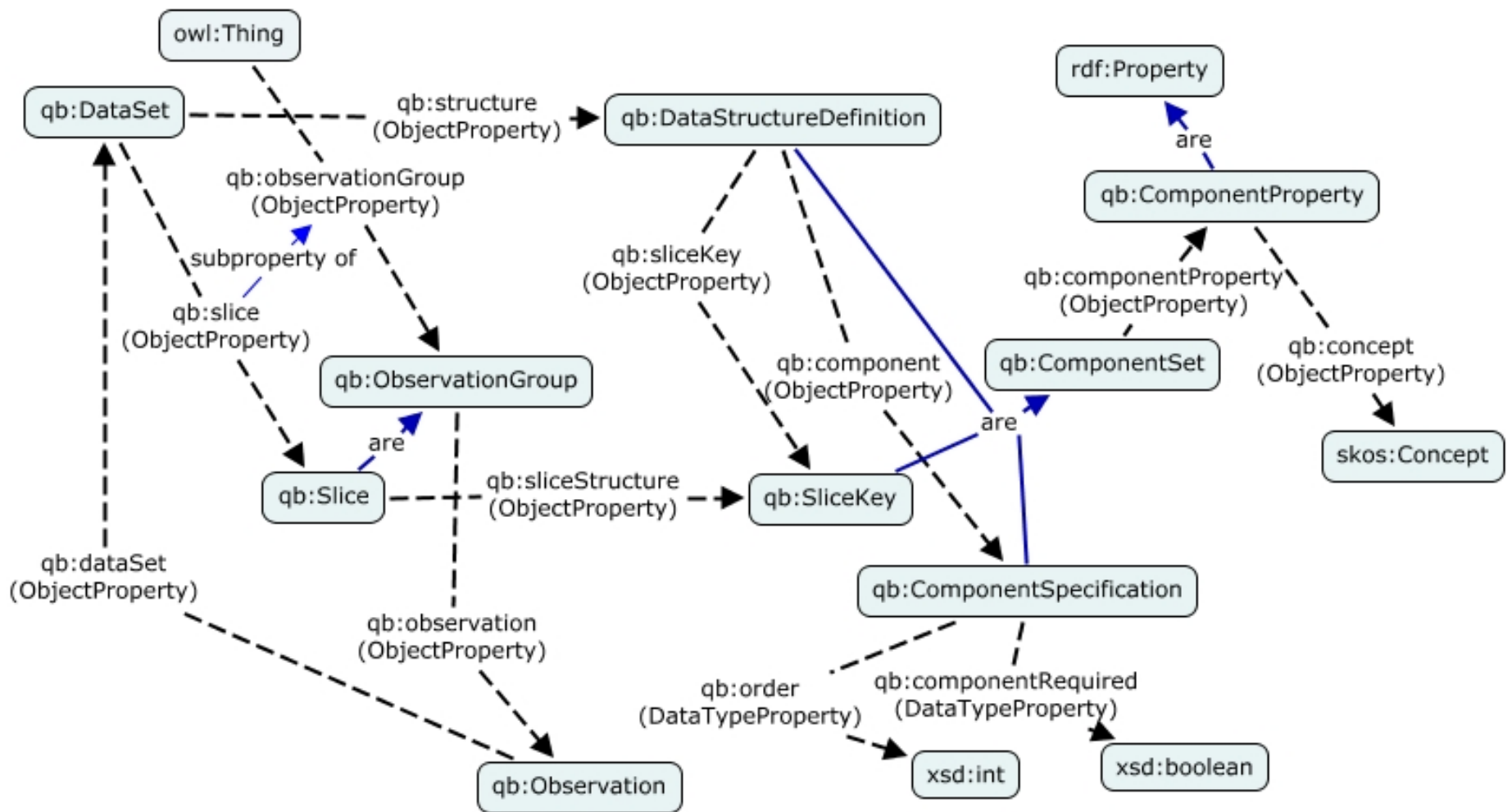
- **Yes!**

- Transition from monolithic tree structure to multi-dimensional data cube → RDF Data Cube Vocabulary (and associated SDMX best practices for slicing it)
- Complex study structure plus user-defined data types → DDI-RDF Discovery (and associated DDI Best practices)

RDF Data cube <http://purl.org/linked-data/cube>

- [RDF Data Cube](#) (qb): a method to organise linked data in slices
 - A vocabulary published by the W3C [Government Linked Data \(GLD\) Working Group](#) (Working Draft)
 - Also the method used to publish statistics data and environmental data in Europe e.g. for Bathing Water Quality in UK <http://www.epimorphics.com/web/projects/bathing-water-quality>
- Advantages
 - Allows multiple views on the same data (similar to OLAP)
 - Generic approach which supports the links to domain-specific definitions
- Useable:
 - In any browser via Linked Data API (HTML output)
 - In JavaScript via Linked Data API (JSON output)
 - In R via SPARQL

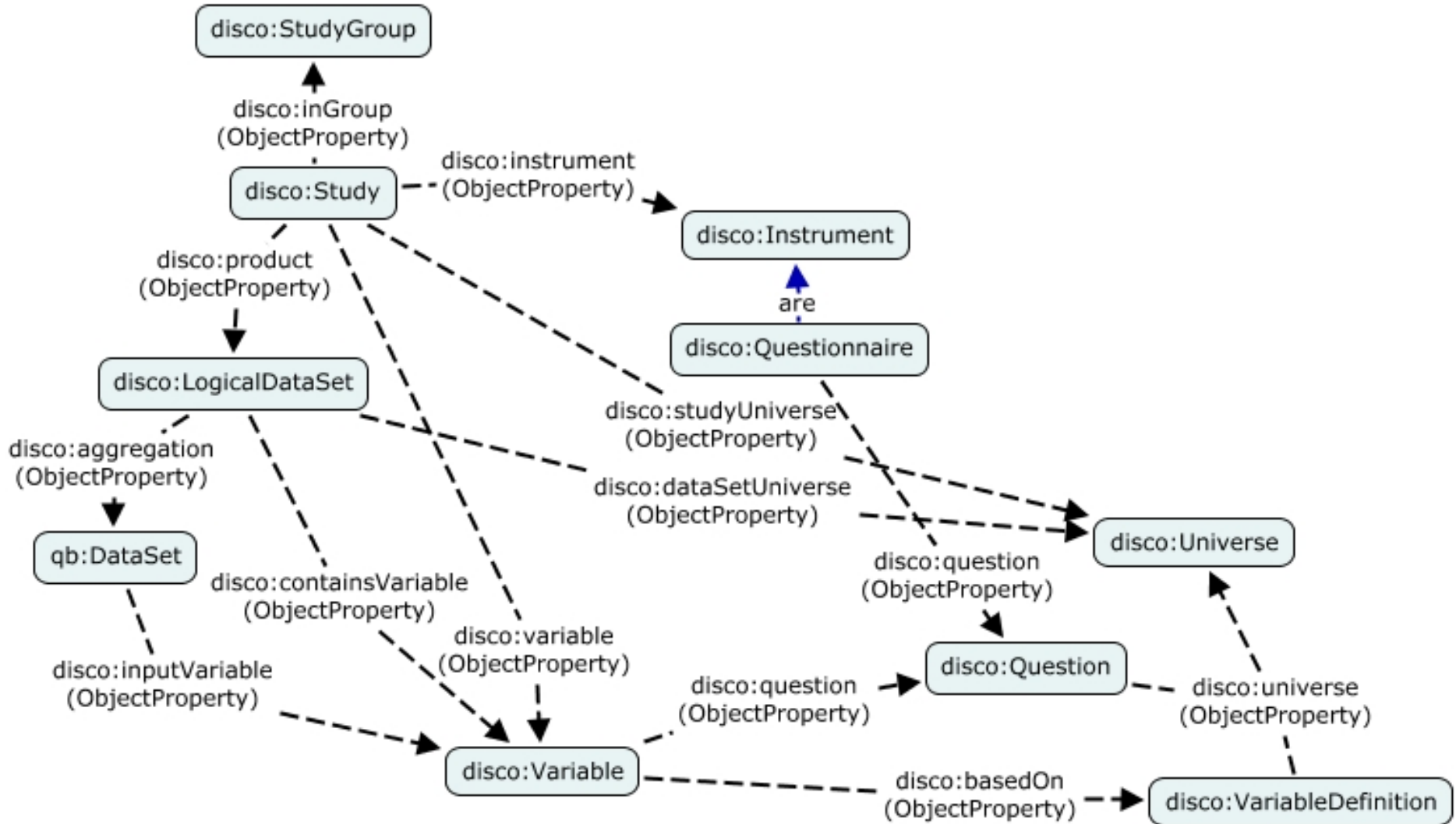
QB (Candidate Rec. version)



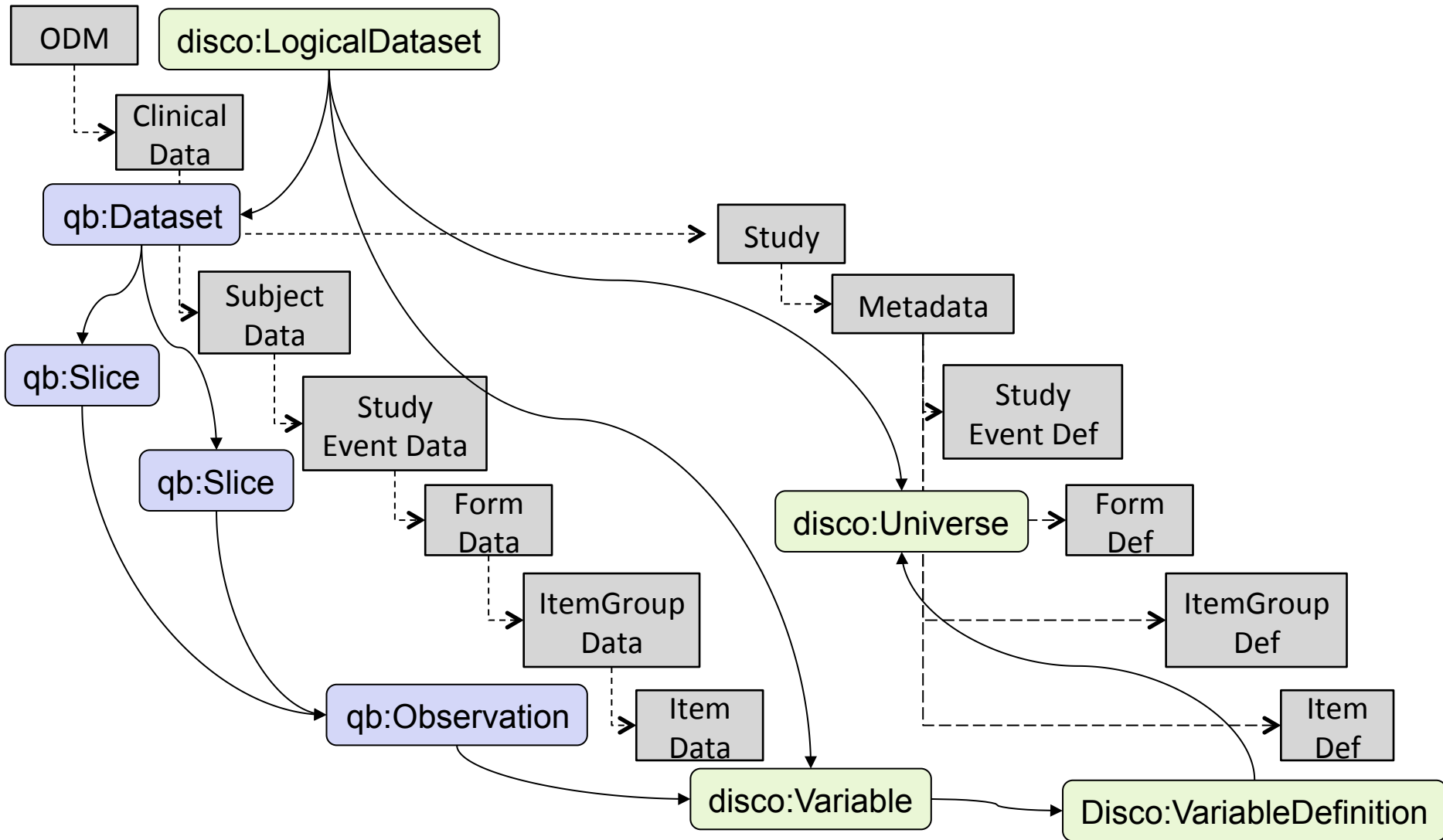
DDI-RDF Discovery

- [DDI-RDF Discovery Vocabulary](#) (disco): a metadata vocabulary for documenting research and survey data
 - [Derived](#) from the Data Documentation Initiative standards
 - DDI and W3C people
 - two Dagstuhl Seminars
 - Designed as a complement to existing vocabularies developed by W3C community (Dublin Core, SKOS, XKOS, DCAT (data catalogue), RDF Data Cube)
- Work in progress
- At least half of it not discussed in this talk - Dataset statistics
 - Useful if we want to attach statistical data to slices ...

Disco (study description)



QB, Disco and ODM



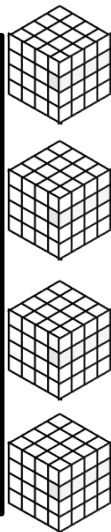
Is this a Semantic Stats problem?

Problem – **Solution** – Remaining challenges

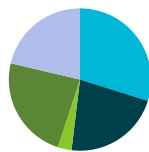
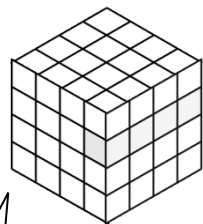
- But ...
 - We have more than one data cube ...
 - 25 after elimination of privacy-sensitive data: patient details, doctor details, ...

Study

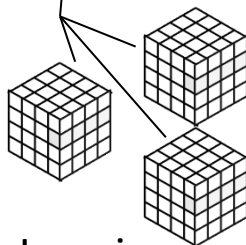
- Screening (Telephone)
- MCI Patient
- AD Patient
- Personal Info
- Assessment Progress



Main cube



1655 variables

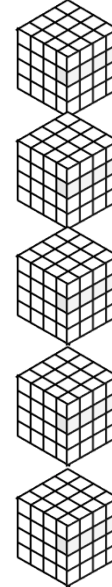


Imaging

- PET -PIB
- MRI
- Dexa

Clinical

- Blood
- CSF
- Vital Signs
- Medical History
- Family History
- Medication



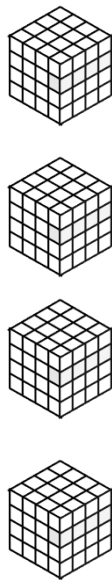
Cognitive

- Neuropsychiatric Inventory Examination
- Neuropsychological Battery
- Memory Complaint Questionnaire
- Short IQ Code



Lifestyle

- Demographics
- Actigraph
- Physical Activity
- Food frequencies
- Food intakes
- Food nutrients
- Alcoholic nutrients



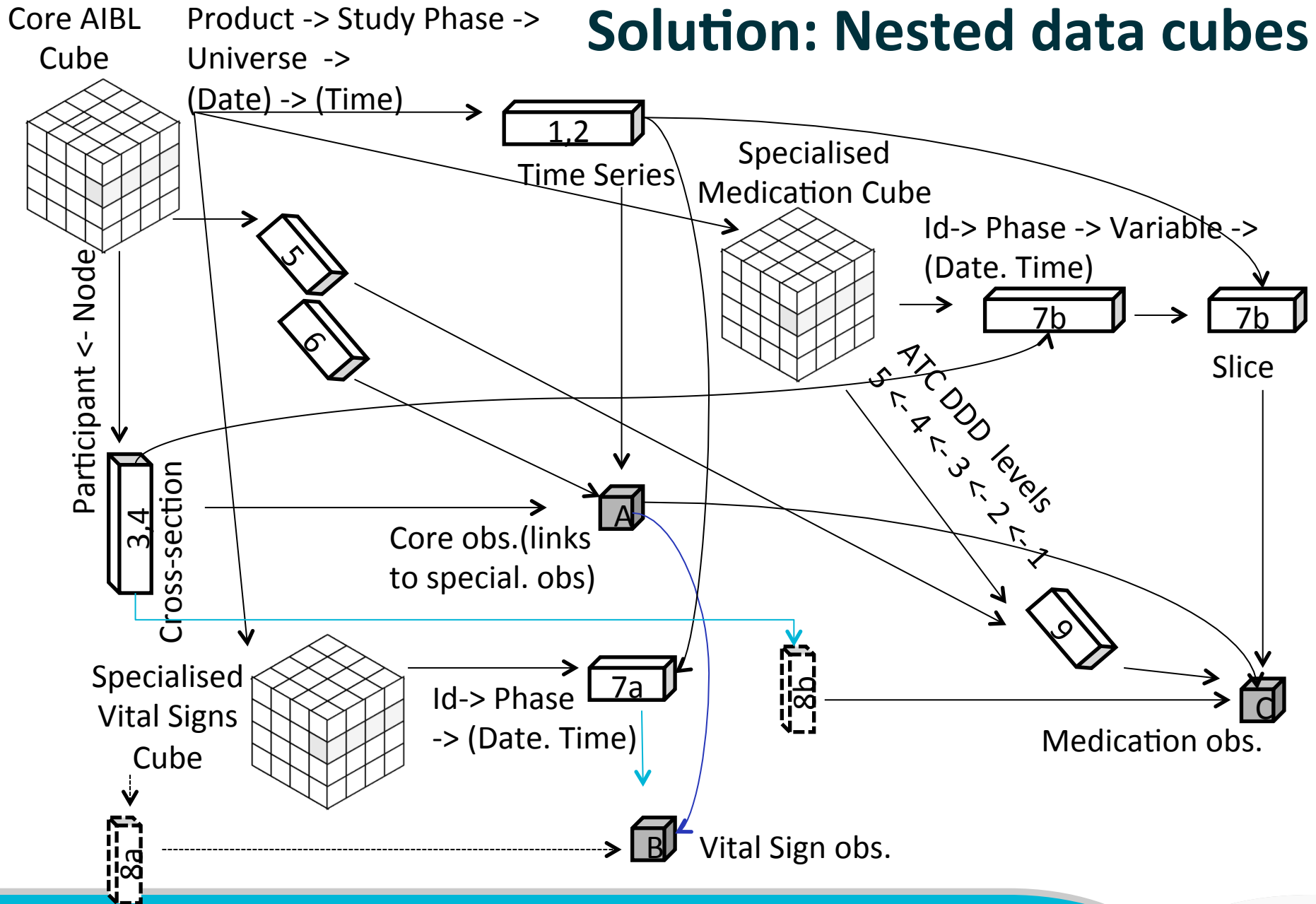
Is this a Semantic Stats problem?

Problem – **Solution** – Remaining challenges

- **Solution: Nested Data Cubes**

- Based on big table defining which variables go in which cubes

Solution: Nested data cubes

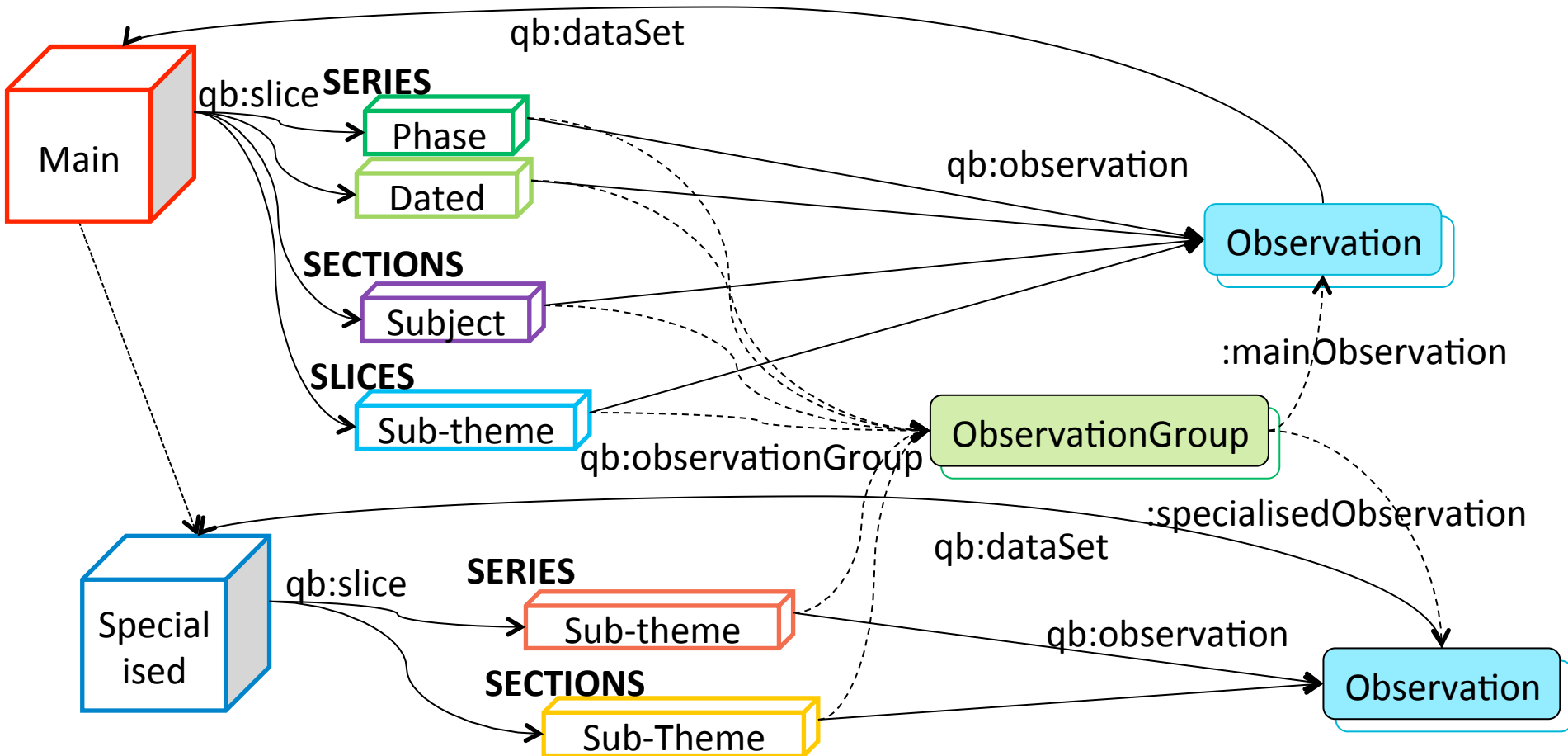


Is this a Semantic Stats problem?

Problem – **Solution** – Remaining challenges

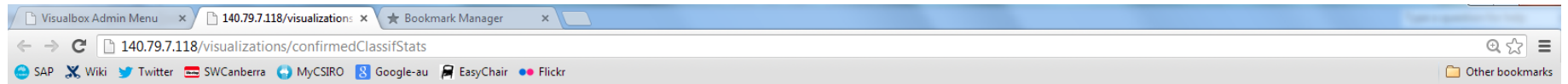
- **Solution: Nested Data Cubes based on RDF Data Cube vocabulary**
 - Maximum compatibility required to be able to reuse tooling developed according to W3C specification
 - Plus URI scheme

Nested Data Cubes with QB

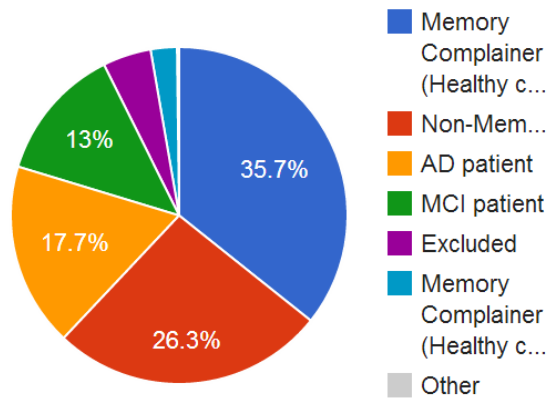


Main cube	URI scheme
Product series (pr)	ROOT/{dataset}/ts/pr/{pr}
Phase series (ph)	ROOT/{dataset}/ts/pr/{pr}/ph/{ph}
Dated series (dt)	ROOT/{dataset}/ts/pr/{pr}/ph/{ph}/dt/{dt}
Product section (pr)	ROOT/{dataset}/cs/pr/{pr}
Node section (nd)	ROOT/{dataset}/cs/pr/{pr}/nd/{nd}
Gender section (gd)	ROOT/{dataset}/cs/pr/{pr}/gd/{gd}
Subject section (su)	ROOT/{dataset}/cs/pr/{pr}/nd/{nd}/su/{su}
Product slices (pr)	ROOT/{dataset}/ds/pr/{pr}
Theme slices (th)	ROOT/{dataset}/ds/pr/{pr}/th/{th}
Sub-theme slices (st)	ROOT/{dataset}/ds/pr/{pr}/th/{th}/st/{st}
Observation groups	ROOT/{dataset}/pr/{pr}/ph/{ph}/su/{su}

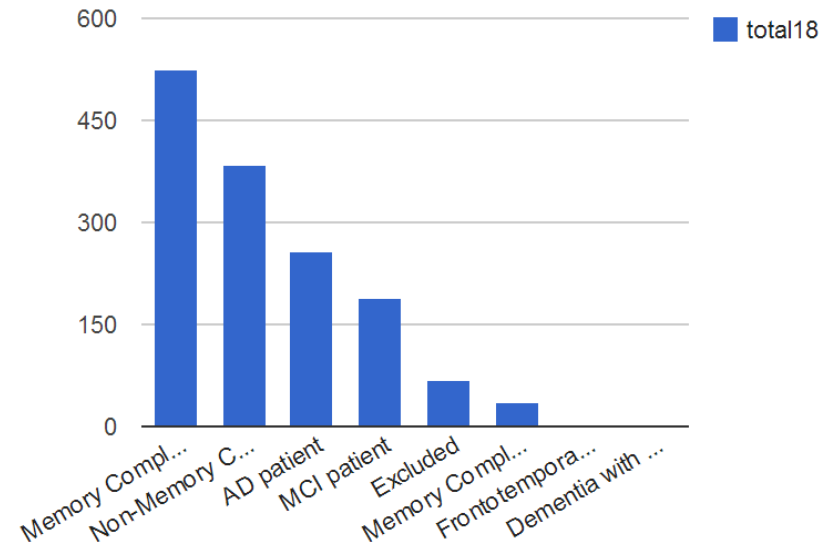
Access via SPARQL (+ Visual Box)



Classification at 18 Months (1)



Classification at 18 Months (2)



Holes

Problem – Solution – Remaining challenges

- Survey-originated data more likely to have missing data
 - Rule-based forms: *if X ... then display box to enter the value of Y*
 - Cases where the patient is no longer available for the study (deceased patient or “lost” patient)
- Question (specs writer): how can you handle datasets with holes?
 - QB example: issue with some IC constraints (see paper)
- Question (tools developers): can you handle datasets with holes (and help the user to avoid them and understand them)?

Sensitive data

Problem – Solution – Remaining challenges

- Need to answer privacy issue with data like AIBL
- Proposal: add different identifiers for each specialised data cube with links at the slice level?
 - To allow browsing /exploration of one data cube at a time with lighter research approval regime (to give enough information for specialist working on data quality issues and for researchers to decide if they need to apply to be granted full access).
 - Question: implementation /access control and performance constraints for queries accessing multiple cubes

Conclusions

Problem – Solution – **Remaining challenges**

- Benefits of semantic statistics vocabularies
- Adoption of SDMX best practices (SDMX guidelines for DSDs Statistical Data and Metadata Exchange 2012)
- Modularity
- Approach reusable for other domains

- Ongoing work on vocabulary mappings (Medications)

- Adoption by Pharma community (FDA/PhUSE)?

Thank you

CSIRO COMPUTATIONAL INFORMATICS

Laurent Lefort
Ontologist

t +61 2 9123 4567

e laurent.lefort@csiro.au

CSIRO COMPUTATIONAL INFORMATICS

www.csiro.au

