**JOURNAL OF
BIOMEDICAL SEMANTICS**

RESEARCH ARTICLE   **Open Access**

Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies

Hugo Leroux[1*†] and Laurent Lefort[2†]

Abstract

Background: There is an increasing recognition of the need for the data capture phase of clinical studies to be improved and for more effective sharing of clinical data. The Health Care and Life Sciences community has embraced semantic technologies to facilitate the integration of health data from electronic health records, clinical studies and pharmaceutical research. This paper explores the integration of clinical study data exchange standards and semantic statistic vocabularies to deliver clinical data as linked data in a format that is easier to enrich with links to complementary data sources and consume by a broad user base.

Methods: We propose a Linked Clinical Data Cube (LCDC), which combines the strength of the RDF Data Cube and DDI-RDF vocabulary to enrich clinical data based on the CDISC standards. The CDISC standards provide the mechanisms for the data to be standardised, made more accessible and accountable whereas the RDF Data Cube and DDI-RDF vocabularies provide novel approaches to managing large volumes of heterogeneous linked data resources.

Results: We validate our approach using a large-scale longitudinal clinical study into neurodegenerative diseases. This dataset, comprising more than 1600 variables clustered in 25 different sub-domains, has been fully converted into RDF forming one main data cube and one specialised cube for each sub-domain. One sub-domain, the Medications specialised cube, has been linked to relevant external vocabularies, such as the Australian Medicines Terminology and the ATC DDD taxonomy and DrugBank terminology. This provides new dimensions on which to query the data that promote the exploration of drug-drug and drug-disease interactions.

Conclusions: This implementation highlights the effectiveness of the association of the semantic statistics vocabularies for the publication of large heterogeneous data sets as linked data and the integration of the semantic statistics vocabularies with the CDISC standards. In particular, it demonstrates the potential of the two vocabularies in overcoming the monolithic nature of the underlying model and improving the navigation and querying of the data from multiple angles to support richer data analysis of clinical study data. The forecasted benefits are more efficient use of clinicians' time and the potential to facilitate cross-study analysis.

Keywords: Ontology, Semantic enrichment, Longitudinal clinical study, RDF data cube, Medication mapping

*Correspondence: hugo.leroux@csiro.au
†Equal contributors
[1]The Australian e-Health Research Centre, Digital Productivity Flagship, CSIRO, Level 5 - UQ Health Sciences Building 901/16, Brisbane, Queensland 4029, Australia
Full list of author information is available at the end of the article

Background

In the last decade, the Health Care and Life Sciences community and pharmaceutical industry have wholeheartedly adopted [1] clinical study data exchange technologies based on XML to capture clinical study data. This is largely due to the recent strategy [2] of the Food and Drug Administration (FDA) in promoting the Clinical Data Interchange Standards Consortium (CDISC) suite of standards to facilitate data submission and exchange. Furthermore, the move by EU and US regulating bodies to open access to clinical data [3,4] will also foster the adoption of tools supporting clinical data management standards, especially those that can easily be linked to methods and tools developed for Government Linked Data and Linked Science Data.

CDISC has developed a set of platform-independent data standards [5] for the collection and dissemination of clinical trial data. The CDISC Operational Data Model (ODM) is an XML format that facilitates the exchange of clinical data captured during a clinical study. ODM-based files contain the study data and the associated descriptions of the data items, their groupings into Case Report Forms (CRFs), which are electronic documents to record the study data, and the associated questions and code lists. Furthermore, the FDA has mandated the use of other CDISC standards in clinical studies. In particular the CDISC Study Data Tabulation Model (SDTM) is used to facilitate study metadata submissions and improve the accountability of the study data. The role of the CDISC Clinical Data Acquisition Standards Harmonization (CDASH) is to standardise the generation of CRFs for clinical studies. The implementation of the ODM, STDM and CDASH standards in Clinical Data Management Systems (CDMS) has enabled larger and more diverse longitudinal clinical research studies and increased the capability of users to exchange and combine data [6].

Challenges relating to the cross-study analysis of clinical study data

A number of limitations relating to the reporting of results derived from current clinical trial endeavours were identified by van Valkenhoef et al. [7]. In particular, they stress that: "*current infrastructure is focused on text-based reports of single studies, whereas efficient evidence-based medicine requires the automated integration of multiple clinical trials from different information resources*" [7]. They specifically advocate for a comprehensive record of clinical trials to be made available in a machine understandable format that would improve the efficiency of evidence-based decision making but more importantly that decisions could then finally be explicitly linked back to the underlying data. Chief among their list of topics for future research directions are: (i) *the development of a comprehensive data model for clinical trials and their aggregate level results*; and (ii) *the development of a platform to share structured systematic review data sets.*

Our contribution: semantic enrichment

This research builds upon existing work [8] to semantically enrich longitudinal clinical study data, based on the CDISC standards, using semantic statistic vocabularies, namely the RDF Data Cube and DDI-RDF vocabularies. We propose a Linked Clinical Data Cube, a set of modular data cubes that helps manage the multi-dimensional and multi-disciplinary nature of the clinical data. The RDF Data Cube vocabulary [9] is used to build multi-dimensional data cubes and supports flexible access to the data via thematic slices. The DDI-RDF Discovery vocabulary [10] is effective at encoding the study-specific data dictionary embedded in the CDISC ODM standard as linked data and helps in managing the link between the data cube variables and the data.

Our objective is to make the data captured within the Australian, Imaging, Biomarker and Lifestyle study of Ageing (AIBL) [11] seamlessly available to researchers who wish to engage in cross-domain analysis of the data. We achieve our goal by semantically enriching the data, when possible, with external data sources. Our approach is four-fold:

Phase 1: Integrating the CDISC ODM data model with the semantic statistic vocabularies. We describe how the clinical data available in CDISC ODM can be mapped to the RDF Data Cube and DDI-RDF Discovery vocabulary to form the Linked Clinical Data Cube.

Phase 2: Splitting the data into modularised cubes. We outline the design principles of splitting the data into more modularised and manageable groupings to provide alternative mechanisms for accessing and querying the data. The RDF Data Cube and DDI-RDF vocabularies are pivotal elements of our slicing strategy and of the URI scheme defined for our implementation.

Phase 3: Enriching the LCDC with the CDISC standards. We discuss how useful the benefits of clinical study data to adhering to the CDISC CDASH and SDTM standards then elaborate on guidelines to classify the data into the broad categories.

Phase 4: Mapping the data to drug terminologies. We demonstrate the utility of the LCDC by mapping the medications data derived from the AIBL study to selected online drug terminologies.

The AIBL study

AIBL is a prospective study of a large group (1112) of individuals residing in two Australian cities, Perth and Melbourne, aged over 60 years who are either classified as cognitively healthy, or meet clinical criteria for mild cognitive impairment or Alzheimer's Disease

and who have agreed to reassessment every 18 months. Assessment comprises extensive study of cognitive function, neuroimaging, blood biomarkers and lifestyle (diet and exercise) characteristics [11]. By combining these investigations in a prospective fashion, the AIBL study contributes to understanding the development and progression of Alzheimer's Disease through the prodromal, preclinical and clinical stages of the disease [12]. It is vital for the clinical data to be reported at regular intervals as the study progresses. To facilitate this task, the study data is manually entered into the OpenClinica Clinical Data Management System (CDMS) by study staff [13]. Figure 1 describes the AIBL study with the five main categories and sub-categories.

OpenClinica [14] is an open-source CDMS for collecting and managing clinical data. The AIBL study data was successfully migrated to this platform in 2011 [13] and has been live since August 2011. OpenClinica supports the creation of customisable studies and the design of user-defined Case Report Forms (CRFs) using an Excel spreadsheet and adheres to the CDISC ODM standard. The data collected for the AIBL study covers multiple domains as shown in Figure 1. This dataset comprises more than 1600 variables clustered into 25 different sub-domains. The study has been split into five themes: *Study*, *Clinical*, *Cognitive*, *Imaging* and *Lifestyle*. The *Study* theme comprises administrative information that, for the most part, is not shared within the cube. Table 1 depicts the total number of instances for the various LCDC classes organised by theme. The total number of variables, in the table, is smaller than 1600 because the generation to RDF suppresses duplicates.

Article outline

In the remainder of this article, we outline an approach to semantically enrich clinical study data, in particular patient-reported medication usage, and facilitate their delivery to clinical researchers. In particular, we outline how the use of semantic statistics vocabularies is effective at organising the data into a LCDC. We also elaborate on the approach taken to categorise the AIBL data set into CDISC CDASH and SDTM domains and the work carried out to translate the CDISC standards into RDF. This leads into the discussion on the design principles for the LCDC and of the benefits of splitting the data into more modularised groupings.

Methods

The LCDC [15] comprises one main cube and several specialised cubes, one for each domain within the study, that integrates the CDISC ODM data model with the RDF Data Cube and DDI-RDF vocabularies. We elaborate further on the rationale behind this integration below. The LCDC is designed around a set of cubes, slices, observation groups and observations and these are discussed further below. The ability to standardise the clinical data in order to facilitate cross-domain and, possibly, cross-study analysis of the data is one of the salient objectives of the LCDC. To this end, we describe how the study variables have been enriched by the CDISC CDASH and SDTM standards. Aside from providing a standardised representation to the study variables and grouping them along the various CDISC categories, this enrichment process allows for seamless substitution of variable names in the navigation and querying of the clinical study. Finally,
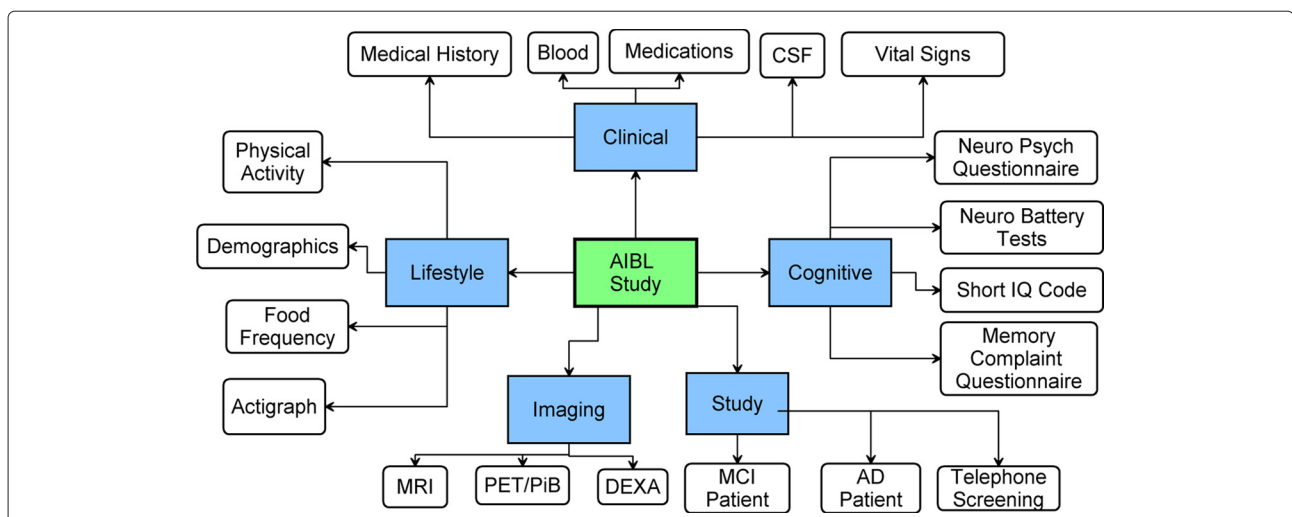

Figure 1 The Australian imaging biomarker and lifestyle study of ageing. Illustrates the logical organisation of the *AIBL* study. The *AIBL* study (depicted as a rectangle in light green with thick border) is split into the five `domains` (depicted as rectangles in light blue). Each domain is further categorised into `sub-domains` depicted by rounded rectangles.

Table 1 Number of instances for the LCDC classes organised by theme

Theme	Total	Obs. group	Obs.	Subject section	Sub-theme series	Sub-theme slice	Variable
Clinical	1030430	4495	25210	1416	25	6452	506
Cognitive	761650	4612	9826	1415	19	4069	367
Imaging	58601	866	2136	365	12	941	59
Lifestyle	710594	4026	11953	1415	19	7360	391
Study	235566	5384	6218	1414	13	3292	155

we outline how the coupling of the study data with external resources - in this case drug terminologies - can be achieved within the LCDC and we elaborate on our process to implement a *linked medications data set* and how the patient-reported medication intake from the AIBL study has been mapped to this data set.

Phase 1: Integrating the CDISC ODM data model with the semantic statistic vocabularies

Clinical study data is extracted in CDISC ODM format. The primary dimensions of the CDISC ODM data model are the Subject and Study Event of interest within the study. The additional dimensions, including the Study, Form, ItemGroup and Item, depend on the study domains and are specified by the data dictionary that defines the study. The strength of the RDF Data Cube is that the original structure of the CDISC ODM data model (Study-Subject-StudyEvent-Form-ItemGroup-Item) lends itself to be replicated in the generated cube with relative ease. A further contribution of the RDF Data Cube is that it can help overcome the monolithic nature of the ODM data model by facilitating the construction of multi-dimensional cubes that offer access points to the data via thematic slices. The LCDC is organised into one main cube and several specialised cubes corresponding to the various domains in the study.

The RDF Data Cube model facilitates the grouping of subsets of observations, within the dataset, whereby all but one (or a small subset) of the dimensions are fixed. Furthermore, it supports alternative methods of accessing the data where the data is aggregated along other dimensions or along the same dimension in different order. The DDI-RDF Discovery vocabulary is used to consistently manage the study-specific data dictionary exported in CDISC ODM format enriched with CDISC metadata resources (CDASH and SDTM). These two vocabularies are supplemented by the Vocabulary of Interlinked Dataset (VoID). These allow the LCDC ontology to be defined with more generalised classes and properties, such as the `disco:Universe`, `disco:Variable` and `disco:VariableDefinition` [15] as depicted in Figure 2.
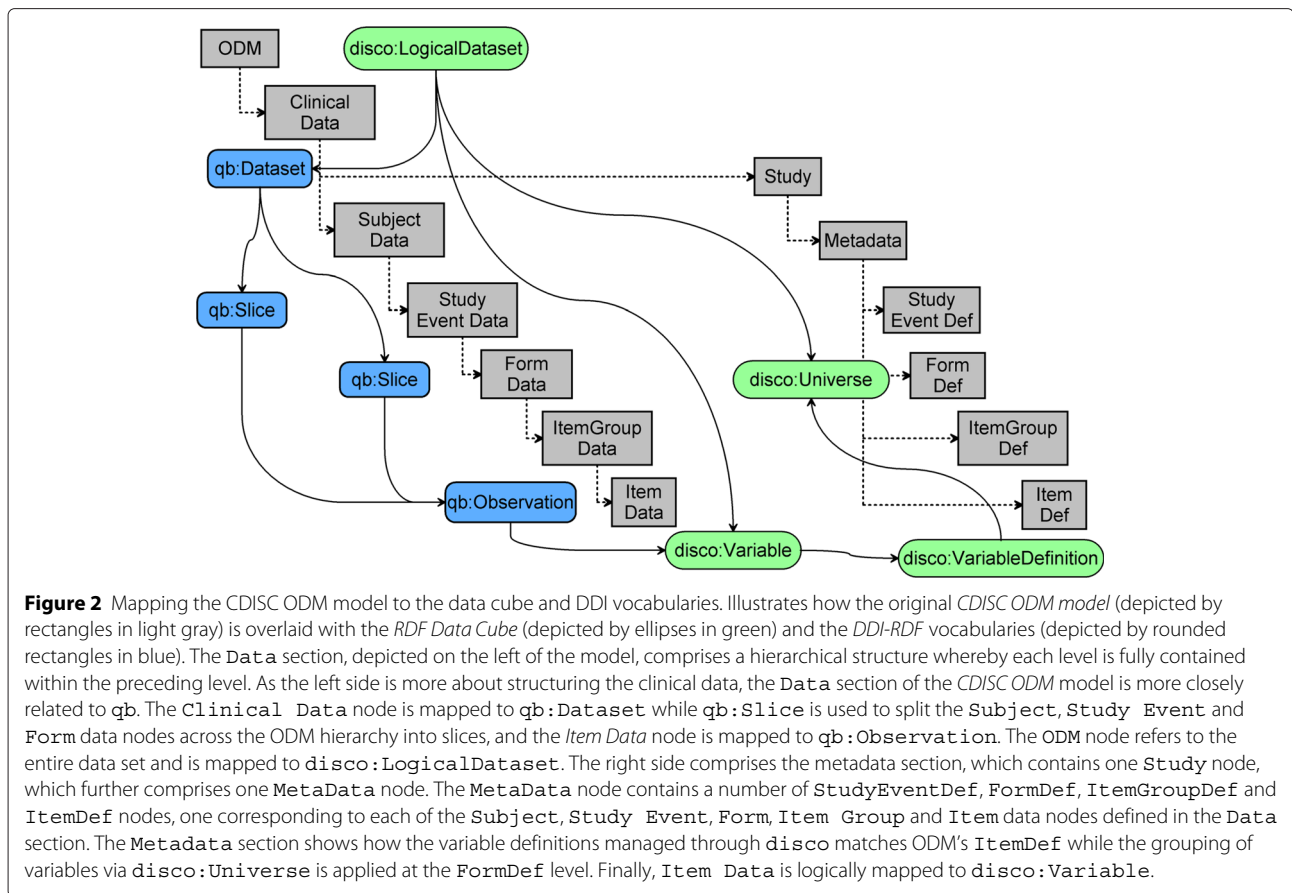
Phase 2: Splitting the data into modularised cubes

The design of the LCDC is achieved in three steps. The first step involves splitting the dataset into smaller, more manageable specialised cubes. The second step is to define several slice hierarchies that offer multiple access options to the individual data records. The third step is to define a URI scheme that supports access to the cube at all levels of the slice hierarchy. These three steps are discussed below.

The LCDC defines three categories of slices. The *time-series* slices address the longitudinal nature of the study and organise the data into time-intervals and dated and non-dated time points. *Cross-section* slices adopt a subject-centric approach to the abstraction of the data along some important concepts such as gender, genotype and neurological classifications. The *Theme* slices categorise the data into the study domains and sub-domains (`disco:Universe` in DDI-RDF) and help link the main and specialised cubes together. This process enhances the navigation and querying of the data in the LCDC because we provide three direct links to nodes containing the data instead of one through the Phase series (at the level of the Study Event data in ODM), the Subject section (at the Subject level) and the sub-theme slice (at the Item Group level).

The slice hierarchy is provided primarily through the use of the classes and properties from the RDF Data Cube. Figure 3 highlights the LCDC slices that subsume `qb:Slice`. We use the `void:subset` property to describe the link between the main and specialised cubes. Links between slices and observations are specified using the `qb:observation` property, while the ones between slices and observation groups are represented by `qb:observationGroup`. The `specialisedSeries` and `specialisedSection` properties manage the links between the slices in the main and specialised cubes. The `specialisedObservation` property, which is a sub-property of `qb:observation`, handles the links between the observation groups from the main cube to corresponding observations in the specialised cubes.

The URI scheme describing the LCDC follows the convention from the Linked Data API [16], which uses URIs ending with an identifier to provide access to a single instance (*Item* endpoint) and URIs ending with

Figure 2 Mapping the CDISC ODM model to the data cube and DDI vocabularies. Illustrates how the original *CDISC ODM model* (depicted by rectangles in light gray) is overlaid with the *RDF Data Cube* (depicted by ellipses in green) and the *DDI-RDF* vocabularies (depicted by rounded rectangles in blue). The `Data` section, depicted on the left of the model, comprises a hierarchical structure whereby each level is fully contained within the preceding level. As the left side is more about structuring the clinical data, the `Data` section of the *CDISC ODM* model is more closely related to `qb`. The `Clinical Data` node is mapped to `qb:Dataset` while `qb:Slice` is used to split the `Subject`, `Study Event` and `Form` data nodes across the ODM hierarchy into slices, and the *Item Data* node is mapped to `qb:Observation`. The `ODM` node refers to the entire data set and is mapped to `disco:LogicalDataset`. The right side comprises the metadata section, which contains one `Study` node, which further comprises one `MetaData` node. The `MetaData` node contains a number of `StudyEventDef`, `FormDef`, `ItemGroupDef` and `ItemDef` nodes, one corresponding to each of the `Subject`, `Study Event`, `Form`, `Item Group` and `Item` data nodes defined in the `Data` section. The `Metadata` section shows how the variable definitions managed through `disco` matches ODM's `ItemDef` while the grouping of variables via `disco:Universe` is applied at the `FormDef` level. Finally, `Item Data` is logically mapped to `disco:Variable`.

a keyword to provide access to a list of instances (*List* endpoint).

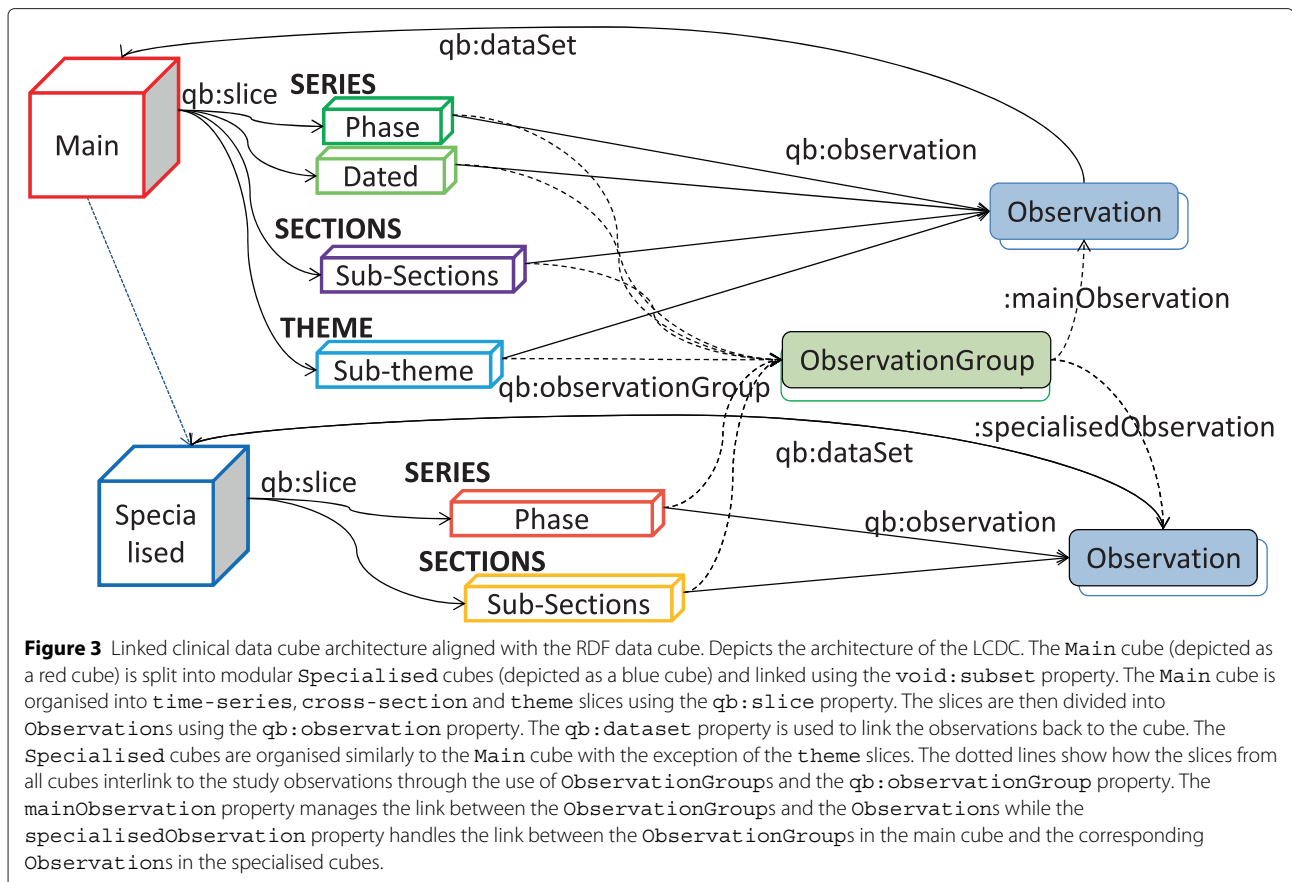**Phase 3: Enriching the LCDC with the CDISC standards**
The CDISC CDASH and SDTM standards provide the means to standardise the clinical data. Despite not being designed around the CDISC standards, there is a good overlap between the AIBL study and the CDISC CDASH and SDTM standards for categories such as Vital Signs, Blood (represented by Laboratory Test in CDASH) and Medical History. For some categories within AIBL, the study data is clustered across many classes that do not necessarily fit to single CDASH or SDTM categories. We have chosen to map our medication data to the Concomitant Medications (CM) class within CDASH. Regarding CM, the approach taken by CDISC is to provide a framework and allow the users the ability to define the terminology of their choice. The AIBL Demographics data can be mapped to the CDISC Demographics and Subject Characteristics categories. SDTM's Trial Arms, Trial Summary, Trial Visits and Subject Visits categories are appropriate targets for mapping longitudinal aspects of the study. For data items that are based on questionnaires, the methodology adopted by CDISC is to guide the user by providing a Questionnaire Supplements (QS) template that the user

can mould to their needs. The SDTM standard provides approximately 50 questionnaires within the QS model that the user can use to model their study. The relatively low number of publicly available questionnaires is due to the fact that many of the questionnaires in clinical studies are licensed.

We have coupled the AIBL-specific variables to existing CDISC concepts, when possible, to allow a straightforward swap of variable names in a query. For example, the AIBL property for systolic blood Pressure (`aiblvitalsigns:systolicBP`) has been linked to the CDISC Vital Sign concept (`cdiscvs:systolicBloodPressure`).

Phase 4: Mapping the data to drug terminologies
In addition to the direct coupling between AIBL and CDISC definitions described above, we have mapped the patient-reported medication intake of the AIBL participants to three external terminologies: AMT, ATC DDD[a] and DrugBank. Our goal is to provide multiple links to hierarchical classifications of drugs. AMT provides unique codes and accurate standardised names to unambiguously identify all commonly used medicines in Australia with eight key top-level concepts [17]. We augment AMT's capabilities with links to ARTG[b] and UNII[c].

Figure 3 Linked clinical data cube architecture aligned with the RDF data cube. Depicts the architecture of the LCDC. The `Main` cube (depicted as a red cube) is split into modular `Specialised` cubes (depicted as a blue cube) and linked using the `void:subset` property. The `Main` cube is organised into `time-series`, `cross-section` and `theme` slices using the `qb:slice` property. The slices are then divided into `Observations` using the `qb:observation` property. The `qb:dataset` property is used to link the observations back to the cube. The `Specialised` cubes are organised similarly to the `Main` cube with the exception of the `theme` slices. The dotted lines show how the slices from all cubes interlink to the study observations through the use of `ObservationGroups` and the `qb:observationGroup` property. The `mainObservation` property manages the link between the `ObservationGroups` and the `Observations` while the `specialisedObservation` property handles the link between the `ObservationGroups` in the main cube and the corresponding `Observations` in the specialised cubes.

ARTG contains the most comprehensive list of brand names (Trade Product) in Australia, while UNII provides a non-proprietary, unambiguous and unique list of substances as maintained by the FDA. DrugBank provides a rich taxonomy of drug information alongside comprehensive drug, gene and food interactions. The appeal for our project is in the exploration of drug-drug interactions that provide some insight into the potential risks and contraindications associated with the intake of the medication. Furthermore, by exploiting the gene-drug interactions of medication targets, we can extend our framework to support the discovery of biomarkers. Finally, the availability of the food interactions will be useful when we explore the association between the participant's drug intake and type and amount of food consumed. Both ATC DDD and DrugBank provide a supplementary means to query the data. The five-level ATC DDD taxonomy of medications provides an additional mechanism for the data to be categorised and offers the means to aggregate the study data for statistical purposes. This is complementary to what is possible with the help of the vocabularies provided by AMT.

Medication mapping is challenging due to the quality, accuracy and completeness of the information. Previous studies [8,18] have identified numerous inconsistencies

linked to the naming of the medications with a mix of *trade name*, *active ingredients* and *informal name* used to describe the prescribed medications.

The processing pipeline for mapping the medications data to the selected medication terminologies is summarised below. The medication records are extracted from OpenClinica, at the start of the pipeline, as an XML document in CDISC ODM format. A data cleaning process is conducted to manually address the inconsistencies described above. This is followed by four mapping phases. In Phase 1, we attempt a map of the "cleaned" medication names to the Trade Product[d] (TP) concept in AMT. We use the list of brand names compiled by ARTG to assist us in this task. In Phase 2, we try to map the same medications to the Medicinal Product[e] (MP) concept in AMT. We use the DrugBank terms to boost the number of mapped concepts. The third phase attempts a map to the substances (active ingredients) either entered by the participants or contained within the medications recorded. To this end, we use the list provided by UNII or the Medicinal Substance[f] (MS) defined in AMT. In Phase 4, we map the medications to the ATC DDD classification hierarchy by taking advantage of the existing mapping between the various terminologies (e.g. DrugBank and ATC DDD). We have thus compiled a linked medications

data set that links AMT, DrugBank, ARTG, ATC DDD and UNII with one another as depicted in Figure 4.

Results

The result of mapping the AIBL medications data to the medication terminologies is illustrated in Table 2. The first row discloses the total number of medications extracted. The second row represents the mappings to either a Medicinal Product, a Trade Product or a Substance in AMT. The third, fourth and fifth rows provide the mapping count for these AMT concepts individually.

The Linked Clinical Data Cube has been evaluated using the full AIBL data set to demonstrate its potential in formulating queries across the broad spectrum of tests and the categories within the clinical study. While simple queries can be answered using a single data cube, more complex queries need data from several cubes to be available. The clinical data is formalised into RDF prior to being loaded in a Virtuoso triple-store.

SPARQL Queries

To demonstrate the utility of the LCDC, we have devised a set of three questions that are typical of the questions that the AIBL researchers are likely to ask of the study data. We provide, below, a listing of the three queries. However, due to privacy constraints, we have structured our queries so that they only return aggregated counts because we are unable to present the participants unique identifier as part of the results of the queries.

Those SPARQL queries have been chosen in order to demonstrate the breadth and depth of questions that may be asked on the data set. They demonstrate how data from

Table 2 Medications mapping statistics

Mapped	Count	Percentage
Total	7942	100.00%
Medicinal product/trade product/substance	5536	69.71%
Trade product	5518	69.48%
Medicinal product	5266	66.31%
Substance	5382	67.77%

the AIBL study can be effortlessly combined with drug information, for example, in order to facilitate queries that answer questions based on drug classifications. Furthermore, we also demonstrate, through the integration of the AIBL data set with terminologies from the CDISC standards, how the AIBL data set can be queried by using the CDISC standardised terminology rather than the actual test names used by the AIBL study. We believe that these types of queries will drive the cross-study and cross-domain benefits of the linked clinical data approaches such as the LCDC.

Query 1: Using CDISC terms, find the number of participants who have hypertension

Hypertension is defined as having systolic and diastolic blood pressure readings above `140` and `90` respectively (written as `140/90 mm Hg`) [19], most of the time. This query explores the use of the CDISC SDTM controlled terminology to access the diastolic and systolic blood pressure readings for participants in the AIBL study. It allows the user to interchangeably use the variable name from the AIBL study or from CDISC SDTM.
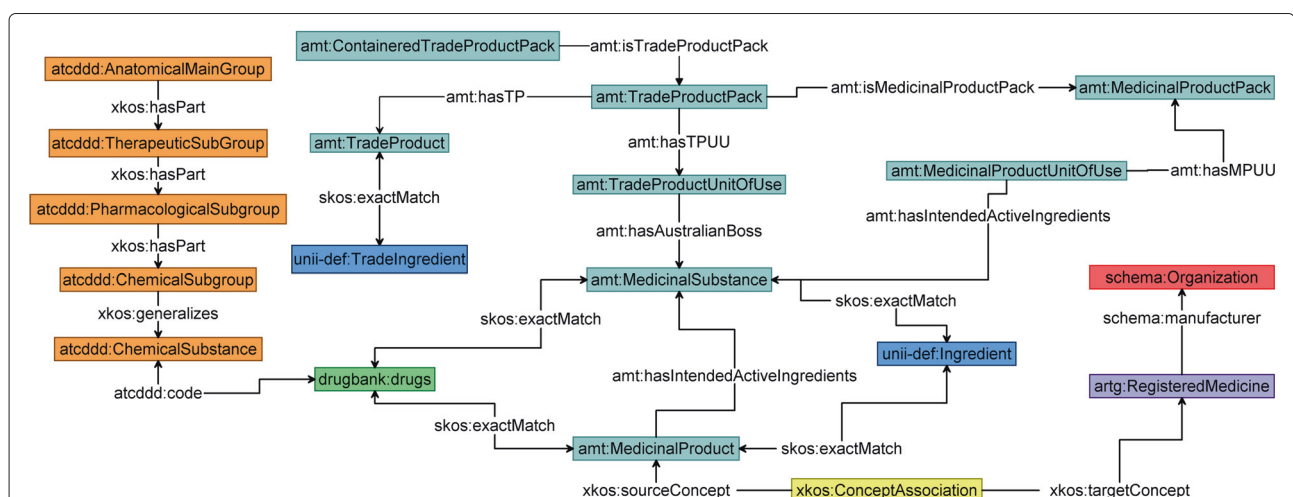

Figure 4 Linked Australian medications data set. Depicts the interlinking of the drug terminologies available, mostly, in Australia in order to facilitate their navigation. For the sake of simplicity, all data item variables have been omitted from the Figure. The `AMT` concepts are depicted in teal. The `ATC DDD` concepts are depicted in orange. `UNII` concepts are in light-blue while `DrugBank` concept is in light green and the `ARTG` concept is in magenta. The Figure also introduces an `xkos:ConceptAssociation` predicate (depicted in yellow) to define many-to-many relationships between `amt:MedicinalProduct` and `artg:RegisteredMedicine` concepts.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lcdcobs: <http://purl.org/sstats/lcdc/def/obs#>
PREFIX cdiscvs: <http://purl.org/odm2rdf/sdtm/vs/def/cdiscvs#>
PREFIX lcdccore: <http://purl.org/sstats/lcdc/def/core#>

SELECT count(DISTINCT ?subject) as ?hypertension WHERE {
     # find the AIBL variable corresponding to cdiscvs:systolicBP
     ?vs_sBP rdfs:subPropertyOf cdiscvs:systolicBloodPressure .
     # find the AIBL variable corresponding to cdiscvs:diastolicBP
     ?vs_dBP rdfs:subPropertyOf cdiscvs:diastolicBloodPressure .
     ?observation rdf:type lcdcobs:Observation .
     #  get the observation for one AIBL subject
     ?observation lcdccore:subject ?subject .
     # get the ?sysBP using cdisc vs alias
     ?observation ?vs_sBP ?sysBP .
     ?observation ?vs_dBP ?diasBP .

     FILTER ((xsd:integer(?sysBP) > 140) && (xsd:integer(?diasBP) > 90) )
}
```

The query obtains the relevant test names from the ontology by performing a lookup of properties that are sub-properties of the CDISC Vital Signs (prefix: vs) diastolic and systolic blood pressure variables. This is achieved by this statement:

```
?vs_dBP rdfs:subPropertyOf cdiscvs:diastolicBloodPressure .
```

This query is possible because we have implemented a *linked set* that connects the variable name from the AIBL study to the standardised terminology in CDISC SDTM vs domain as illustrated below.

```
aiblvitalsigns:diastolicBP
      rdf:type owl:DatatypeProperty ;
      rdfs:subPropertyOf cdiscvs:diastolicBloodPressure.

aiblvitalsigns:systolicBP
      rdf:type owl:DatatypeProperty;
      rdfs:subPropertyOf cdiscvs:systolicBloodPressure.
```

We believe that the use of linksets in this manner is important and useful because it adheres to the principles of *information hiding* in that the user need not be aware of the exact wording of a variable. As long as the user knows the corresponding standardised variable name, the user is able to successfully execute a query on the data set. We intend to further develop this traceability mechanism with the help of the Provenance Ontology [20] to fully disclose how the published data is derived from the originally captured data.

The result of Query 1 is displayed below:

hypertension

242

Query 2: How many participants are taking an anti-diabetic drug such as Metformin?

Some studies [21,22] have shown a possible link between *type2 diabetes* and early-stage AD. In this query, we retrieve a list of anti-diabetic drugs to demonstrate the benefits of linking the patient-reported medications to standardised external terminologies and the strength of the LCDC in using federated queries to facilitate cross-domain querying. The first portion of this query obtains a list of anti-diabetic drugs from DrugBank (outlined in section A in the SPARQL). The second part of the query utilises the mappings between the patient-reported medications and DrugBank entities to link to the anti-diabetic drugs identified in section A.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX drugbank: <http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource
                  /drugbank/>
PREFIX aiblmed: <http://aehrc-ci.it.csiro.au/aibl/lcdc/clinical/medication/def
                  /aibl-medication#>
```

```
PREFIX lcdcobs: <http://purl.org/sstats/lcdc/def/obs#>
PREFIX lcdccore: <http://purl.org/sstats/lcdc/def/core#>
PREFIX cm:       <http://purl.org/sstats/lcdc/cm/def/cm#>
PREFIX amt:      <http://nehta.gov/amt#>

SELECT count (distinct ?subject) as ?count ?mp_med
WHERE {
    # Section A. find all instances of anti-diabetic drugs
    SERVICE <http://wifo5-04.informatik.uni-mannheim.de/drugbank/sparql> {
        # find the drug (?s) that has the name 'Metformin'
        ?s drugbank:genericName "Metformin" .
        # find the category of the drug (?s)
        ?s drugbank:drugCategory ?category .
        # find all other instances of ?drug that has the same ?category
        ?drug drugbank:drugCategory ?category .
     }

    # Section B. find the participants who take the ?drug from A.
    { SELECT distinct ?drug ?med ?subject ?mp_med WHERE {
      GRAPH <http://localhost/dataset/aibl/lcdc/clinical> {
          # specifies that ?obs is an observation
          ?obs a lcdcobs:Observation .
          # get the medicinal product code for this med
          ?obs cm:medicinalProduct ?cm_mp .
          # lookup the drugbank entity linked to this MP
          ?cm_mp skos:exactMatch ?drug .
          # get this drug's name
          ?cm_mp amt:synonym ?mp_med .
          # find the participant associated with this observation
          ?obs lcdccore:subject ?subject .
      }
    } }
}
LIMIT 20
```

The *linkset* developed to map the AMT concepts to DrugBank has been inspired from the approach described in [23,24] and uses the skos:exactMatch predicate.

```
<http://snomedtools.info/snomed/version/1/concept/rdfs/105271000036100>
      rdf:type amt:MedicinalProduct ,
              owl:NamedIndividual ;
      skos:exactMatch
        <http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugs/DB06655>.
```

The significance of this mapping is the provision of *drug-drug*, *drug-gene* and possibly *drug-disease* and *gene-gene* information relating to the AIBL study to the researchers by fully utilising the links provided by DrugBank.

The result of Query 2 is displayed in Table 3:

Query 3: Are there participants whose classification has transitioned from healthy to mild cognitive impairment but whose triglyceride's level has remained normal?

Research has investigated the risk factors associated with *low-density lipoproteins* or *triglycerides* on the incidence and progression of dementia and AD in later life [22]. With this in mind, we construct the query below to retrieve participants' records whose confirmed classification status have been updated from being healthy as subjective memory complainer or non-memory complainer to having mild cognitive impairment but who have also maintained a normal (< 1.7 mmol/L) level of triglycerides in their blood sample over the course of an 18-month period between the baseline and 18-month time-points.

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX lcdcsection: <http://purl.org/sstats/lcdc/def/cross-section#>
PREFIX lcdccore: <http://purl.org/sstats/lcdc/def/core#>
PREFIX aiblblood: <http://aehrc-ci.it.csiro.au/aibl/lcdc/clinical/blood/def
                        /aibl-blood#>
```

```
PREFIX aiblneuropsych: <http://aehrc-ci.it.csiro.au/aibl/lcdc/cognitive/neuropsych
                        /def/aibl-neuropsych#>
PREFIX aiblphase: <http://aehrc-ci.it.csiro.au/dataset/aibl/lcdc/id/phase/>
PREFIX aiblsubtheme: <http://aehrc-ci.it.csiro.au/dataset/aibl/lcdc/id/subtheme/>

SELECT DISTINCT count(?subject) as ?subjectCount ?class1 ?class2
WHERE {
  { select distinct ?subject ?obs1a ?trig1 where {
      # retrieve the objects from the SubTheme cross-section slice
      ?nodeSect a lcdcsection:SubThemeSection .
      # only get observations for the participants from the Blood domain
      ?nodeSect lcdcsection:subtheme aiblsubtheme:blood .
      # get the observations from the slice
      ?nodeSect qb:observation ?obs1a .
      # once we get the observations, get the subject
      ?obs1a lcdccore:subject ?subject .
      # only select observations for the baseline phase
      ?obs1a lcdccore:phase aiblphase:baseline .
      # get the triglycerides measurements
      ?obs1a aiblblood:trig ?trig1 .
  } }
  { select distinct ?subject ?class1 ?obs1b where {
      ?nodeSect a lcdcsection:SubThemeSection .
      # only get observations for the participants from the Neuropsych domain
      ?nodeSect lcdcsection:subtheme aiblsubtheme:neuropsych .
      ?nodeSect qb:observation ?obs1b .
      ?obs1b lcdccore:subject ?subject .
      ?obs1b lcdccore:phase  aiblphase:baseline .
      # get the subject's classifications
      ?obs1b aiblneuropsych:confirmedClassification ?class1 .
      # only select healthy subjects
      FILTER(?class1 =
          aiblneuropsych:ConfirmClassification360-memoryComplainerHealthyControl
               || ?class1 =
          aiblneuropsych:ConfirmClassification360-nonMemoryComplainerHealthyControl
      )
  } }
  FILTER (xsd:float(?trig1) < 1.7)
  { select distinct ?subject ?obs2a ?trig2 where {
      ?nodeSect a lcdcsection:SubThemeSection .
      ?nodeSect lcdcsection:subtheme aiblsubtheme:blood .
      ?nodeSect qb:observation ?obs2a .
      ?obs2a lcdccore:subject ?subject .
      # select observations for the 18-month phase
      ?obs2a lcdccore:phase  aiblphase:18months .
      ?obs2a aiblblood:trig ?trig2 .
  } }
   { select distinct ?subject ?class2 ?obs2b where {
      ?nodeSect a lcdcsection:SubThemeSection .
      ?nodeSect lcdcsection:subtheme aiblsubtheme:neuropsych .
      ?nodeSect qb:observation ?obs2b .
      ?obs2b lcdccore:subject ?subject .
      ?obs2b lcdccore:phase  aiblphase:18months .
      ?obs2b aiblneuropsych:confirmedClassification ?class2 .
      # ensure subject have transitioned to MCI
      FILTER(?class2 = aiblneuropsych:ConfirmClassification360-mciPatient )
  } }
  # normal range for trig. is < 1.7 mmol/L
  FILTER (xsd:float(?trig2) < 1.7)
}
```

The above query highlights the strength of the LCDC in facilitating cross-domain queries by fully exploiting the potential of slices and observations within the specialised cubes. While the above query can be achieved without a data cube, the use of slices and observations make the query more elegant and effective. It demonstrates the navigation of the AIBL data set across two specialised cubes (Neuropsych and Blood) and four slices (two slices at each time points for each cube). These are contained within the four observations (?obs1a, ?obs1b, ?obs2a, ?obs2b) within the above query.

Table 3 Participants taking anti-diabetic drugs

Count	mp_med
1	Insulin glargine
3	Glimepiride
46	Metformin
4	Rosiglitazone
2	Glipizide
17	Gliclazide
4	Pioglitazone
1	Sitagliptin

The result of Query 3 is displayed in Table 4:

We provide an indication of the execution time of the three queries in Table 5 below. These queries have been executed on a Virtuoso 6.1 instance running on a virtual machine with an AMD Opteron Processor 62xx CPU, 8GB of DDR3 RAM and running Ubuntu 13.04 LTS (Raring Ringtail).

Discussion and related work

Our results demonstrate the effectiveness of integrating semantic statistics vocabularies with the CDISC standards in order to expedite the navigation and querying of the data. Our contribution extends previous attempts to semantically enrich biomedical research data using ontologies [25] or linked data resources [26]. To the best of our knowledge, no study has yet investigated the association of semantic statistics vocabularies with clinical data exchange standards. The design of the LCDC was inspired by the Translational Medicine Ontology [27] and our use cases were motivated by similar objectives of providing qualitative and pertinent clinical data to the researchers and clinicians in the right format. This is what has driven our resolve to split the data into one main cube and several specialised cubes corresponding to the various domains in our study. The benefits of this approach are demonstrated in the third query where data from two specialised cubes are amalgamated to derive the results.

Observational clinical study data is patchy by nature, mainly because of the various collection mechanisms involved that often lead to information being inadvertently left out or inaccurately recorded. Furthermore, the sheer volume of variables and the longitudinal nature of the AIBL clinical study have given rise to an enormous volume of data that need to be analysed. This has led to the second design decision that is to split the data into time-series, cross-sections and themes in order to improve their manageability during the generation process and facilitate their discovery and usability by end users. Moreover, the addition of external standardised terminologies, such as the CDISC standard terminologies and the various drug vocabularies utilised, have contributed not only to standardising the data and to removing ambiguities but to enriching the data by providing links to relevant online resources, such as genes and pathways definitions and information about their interactions with the entities.

Challenges in the use of the CDISC standards as the underlying model

While the CDISC models suit our immediate purpose, they present a few shortcomings, mainly in relation to the semantics associated with the clinical study data. ODM's constrained hierarchical structure largely promotes single-study explorations of clinical study data. Furthermore, the inability to store domain information alongside the user-defined data items in the customisable CRFs is, in our view, very restrictive, thus impeding their use outside of the study context [28]. However, this stems more from the various failings in the implementation of the CDISC standards by the vendors. The ODM standard allows for CDASH terms to be inserted through the use of annotations within the ODM XML model. However, several vendors, such as OpenClinica, choose not to offer this feature natively within their tool. Abler et al. [28] make a passionate claim for the definition of a *language of forms* that can effectively record the *logical relationships between questions or sets of questions* asked in the forms.

On a more technical aspect, ODM also suffers from a lack of established complex data type standards, thus allowing a study coordinator to provide an alternative definition for, say, the *Physical Quantity* data type. Furthermore, despite the provision of detailed *Implementation Guides* describing the correct way of encoding data items, the definition of very coarsely granulated meta-data categories, such as *Medical History* in SDTM, opens up the possibility, for the user, to capture semantically identical data in multiple domains. While the lack of data standards is a problem, the lack of mechanisms to enforce adherence to these data standards is a greater problem. As such, despite CDISC providing mechanisms, through its SDTM and CDASH models, to define common semantics, in our experience, very few study coordinators choose to use them.

Our choice of the CDISC standard as the underlying model for our architecture is influenced by three factors: (i) since the FDA and other regulatory bodies mandate the use of CDISC as the *de facto* standards for representing and reporting clinical study data, a vast majority of the clinical study data that we encounter is already in CDISC format; (ii) several extensions to the CDISC standards (such as the *Therapeutic Area* standard for Alzheimer's Disease) are appealing to us; and (iii) we have *not yet* found a consistent and complete set of ontologies that we could use instead.

In our approach to semantically enrich the clinical study data, we need to address the study-specific nature of

Table 4 Participants' classifications and triglycerides level

SubjectCount	Class1	Class2
11	ConfirmClassification360-memoryComplainerHealthyControl	ConfirmClassification360-mciPatient
4	ConfirmClassification360-nonMemoryComplainerHealthyControl	ConfirmClassification360-mciPatient

CDISC ODM datasets. We inherit many issues that have been created in the previous steps of the data capture chain, such as the use of user-defined questionnaires and instruments that use their own language and the loss of domain knowledge during the digitisation phase of the data. Our solution is to reintroduce the loss of domain-specific information by first trying to retrofit the study variables to the SDTM and CDASH models, even though they were not initially modelled that way. Concurrently, we look to biomedical ontologies, such as the NCBO Bioportal ontologies [29] and SNOMED CT[g], to provide alternative foundations for domain enrichment of the data set. Several ontologies, in the context of clinical trials [30-32], have been proposed recently and are partially applicable to our needs. However, they do not adequately cover the observational aspects that are required for our data cubes. Furthermore, several of these ontologies have a large number of dependencies to other ontologies that do not meet our requirements. We overcome the limitations related to the single-study nature of ODM by fitting the study data to the RDF Data Cube. The introduction of additional dimensions, through the integration to the RDF Data Cube, opens up new access points to the data through the use of the thematic slices.

Ultimately, our view is that regulatory bodies have a pivotal role to play in encouraging the clinical study coordinators to engage with data scientists at an earlier stage in their clinical study to help with the design of their study and associated artefacts. Too much emphasis is placed on the data collection phase and not enough effort is expended in clarifying what is needed to analyse the data.

Related work

The Linked Open Drug Data (LODD) [33] and the Linked Life Data (LLD) [34] projects provide additional resources that can be used to extend the Linked Clinical Data Cube. Both projects aim to build a large scale knowledge cloud that can be used for drug discovery. LODD federates the efforts by participants of the W3C Semantic Web Health Care and Life Sciences (HCLS) Interest group to convert available resources into linked data. LLD provides a semantic data integration platform for the biomedical domain comprising many of the data sources belonging to LODD. The resulting datasets contains more than 8 million triples representing the knowledge within over 2 millions links relating to medications, diseases, clinical trials, gene information and pharmaceutical companies among others. This was followed by efforts to convert the ChEMBL database as linked open data [23]. This new linked dataset combines the description of the biological entities with links to *Bio2RDF* [35], *ChemSpider* [36], *OpenMoleculesRDF* [37] and *CrossRef* [38] to allow dereferenceable access to a myriad of external datasets. We have adopted a similar methodology in our approach to map the medications specialised cube to AMT, DrugBank and ATC DDD.

Among the various use cases reported via the W3C HCLS Interest group are efforts to explore links to identify and verify genes linked to Alzheimer's disease (AD). Through the links between the drug, medications, disease and clinical trial repositories, we hope to leverage on efforts by others to further explore the effects of prescribed medications, for AD sufferers, on the various genes comprising the pathways of interest. Other applications of LODD include the identification of potential side-effects linked to the intake of drugs that have conflicting stimuli on the disease pathways.

The SALUS project [39] is a former attempt to adapt CDISC standards to build a Semantic Framework to improve interoperability between clinical research and clinical care domains. We adopt a similar approach to them but their focus is on service mappings rather than linked data sets. The Semantic Cockpit [40] project aims to develop a data slicing framework comparable to what we propose on the basis of the RDF Data Cube. The goal of this project is to intelligently assist business analysts by discriminating unimportant information and using reasoning to only present useful information to the analyst.

Table 5 Query performances

Query	Execution time (msec)
1	22
2	36
3	270

The *Linked Medical Data Access Control* (LiMDAC) project [41] has devised a framework to enable the integration of medical data without compromising its privacy, security and integrity. It defines three linked data models that use the RDF Data Cube to build an access control framework that restricts access to the aggregated data.

The *Pharmaceutical Users Software Exchange* [42] community, in concert with the FDA, has started work on RDF representations of various CDISC models [43], including the terminologies published by the *National Cancer Institute (NCI) Enterprise Vocabulary Services* [44]. This community has started to evaluate the RDF Data Cube [45,46] for the publication of clinical study data. These conversions of comma-separated-value files, however, do not fully exploit the relationships between the data and metadata structures embedded within the XML versions of the CDISC standards and the patterns and concept definitions included in the generated RDF content.

Conclusions

This paper has outlined the semantic enrichment of longitudinal clinical study data based on the CDISC standards with elements from the semantic statistics vocabularies, namely the RDF Data Cube and the DDI-RDF Discovery vocabularies. We have outlined how the Health Care and Life Science community is likely to benefit from the adoption of tools and techniques that will deliver clinical data as linked data and advance its integration with complementary data sources. In this regard, we have proposed a Linked Clinical Data Cube, which integrates one main and several specialised data cubes to provide increased flexibility in the navigation of the clinical data and allow the users to formulate the queries more efficiently and effectively. The Linked Clinical Data Cube combines the strength of the RDF Data Cube in defining multi-dimensional data cubes and the DDI-RDF Discovery vocabulary in encoding the CDISC metadata and the study specific data dictionary as linked data. Our approach was validated using data captured as part of a longitudinal clinical study into neurodegenerative diseases. This research has resulted in four contributions. First, we have uncovered the complementarities of the RDF Data Cube and DDI-RDF Discovery vocabularies for the publication of large heterogeneous data sets as linked data. Second, we have demonstrated the fit of the semantic statistics vocabularies to enrich the CDISC ODM data model for the publication of clinical study data as linked data. Third, we have illustrated how the clinical study data has been semantically enriched with links to external resources and how they ultimately improve the navigation and querying of the data. Fourth, we have built the foundations of a framework supporting cross-domains and cross-study analysis by adopting a more standardised data structure. Our next step is to enrich the remaining study data set with concepts from other domain ontologies, such as Blood, Neuropsychological tests and Nutrition, to name just three.

Endnotes

[a]Anatomical Therapeutic Chemical Defined Daily Dose.
[b]Australian Register of Therapeutic Goods.
[c]Unique Ingredient Identifier.
[d]`30560011000036108 | trade product |.`
[e]`30497011000036103 | medicinal product|.`
[f]`30388011000036105 | medicinal substance |.`
[g]Systematized Nomenclature of Medicine Clinical Terms.

Author details
[1]The Australian e-Health Research Centre, Digital Productivity Flagship, CSIRO, Level 5 - UQ Health Sciences Building 901/16, Brisbane, Queensland 4029, Australia. [2]Digital Economy Program, Digital Productivity Flagship, CSIRO, Canberra, ACT 2601, Australia.

References

1. Lu Z, Su J. Clinical data management: Current status, challenges, and future directions from industry perspectives. Open Access J Clin Trials. 2010;2:93–105.
2. FDA: Data Standards Strategy Version 1.0. 2012. http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM332955.pdf.
3. Eichler H-G, Abadie E, Breckenridge A, Leufkens H, Rasi G. Open clinical trial data for all? A view from regulators. PLoS Med. 2012;9(4):1001202.
4. IOM. Sharing Clinical Research Data - Workshop Summary. The National Academies Press. 2013.9–17. http://www.iom.edu/Reports/2013/Sharing-Clinical-Research-Data.aspx.
5. van Valkenhoef G, Tervonen T, de Brock B, Hillege H. Clinical trials information in drug development and regulation: existing systems and standards. Netherlands: University of Groningen Press; 2012.
6. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Methods Inf Med. 2009;48(1):45–54.
7. van Valkenhoef G, Tervonen T, de Brock B, Hillege H. Deficiencies in the transfer and availability of clinical trials evidence: A review of existing systems and standards. BMC Med Informatics Decision Making. 2012;12(1):95.
8. Leroux H, Lefort L. Using CDISC ODM and the RDF Data Cube for the semantic enrichment of longitudinal clinical trial data. In: Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences. Paris, France: CEUR Workshop Proceedings; 2012.

9. Cyganiak R, Reynolds D, Tennison J. The RDF data cube vocabulary. World Wide Web Consortium. 2014. http://www.w3.org/TR/vocab-data-cube/.

10. Bosch T, Cyganiak R, Zapilko B, Cotton F, Gregory A, Kämpgen B, et al. DDI-RDF Discovery vocabulary a vocabulary for publishing metadata about data sets (research and survey data) into the web of linked data unofficial draft. Technical report, DDI Alliance; June 2013.

11. Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatrics/IPA. 2009;21(4):672–87.

12. Ellis KA, Rainey-Smith SR, Rembach A, Macaulay SL, Villemagne VL. Enabling a multidisciplinary approach to the study of ageing and alzheimer's disease: An update from the Australian Imaging Biomarkers and Lifestyle (aibl) study. Int Rev Psychiatry. 2013;25(6):699–710.

13. Leroux H, McBride S, Gibson S. On selecting a clinical trial management system for large scale, multi-centre, multi-modal clinical research study. Stud Health Technol Informatics. 2011;168:89–95.

14. OpenClinica https://www.openclinica.com/.

15. Lefort L, Leroux H. Design and generation of Linked Clinical Data Cubes. In: Proceedings of 1st International Workshop on Semantic Statistics (SemStats 2013). Sydney, Australia; 2013.

16. Linked Data API https://code.google.com/p/linked-data-api/.

17. Michel J, Lawley MJ, Chu A, Barned J. Mapping the queensland health ipharmacy medication file to the Australian Medicines Terminology using Snapper. Stud Health Technol Informatics. 2011;168:104–16.

18. McBride SJ, Lawley MJ, Leroux H, Gibson S. Using Australian Medicines Terminology (AMT) and SNOMED CT-AU to better support clinical research. Stud Health Technol Informatics. 2012;178:144–9.

19. High Blood Pressure http://www.nlm.nih.gov/medlineplus/highbloodpressure.html.

20. PROV-O: The PROV Ontology http://www.w3.org/TR/prov-o/.

21. Griggs J. Eating your way to dementia. New Scientist. 2013;220(2945):6–7.

22. Kloppenborg RP, van den Berg E, Kappelle LJ, Biessels GJ. Diabetes and other vascular risk factors for dementia: which factor matters most? A systematic review. Eur J Pharmacol. 2008;585(1):97–108.

23. Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, Tkachenko V, et al. The CHEMBL database as linked open data. J Cheminformatics. 2013;5(1):1–12.

24. Groth P, Loizou A, Gray AJ, Goble C, Harland L, Pettifer S. API-centric linked data integration: The open PHACTS discovery platform case study. Web Semantics: Science, Services and Agents on the World Wide Web. 2014;29(0):12–8. http://dx.doi.org/10.1016/j.websem.2014.03.003, http://www.sciencedirect.com/science/article/pii/S1570826814000195.

25. Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the semantic web. Brief Bioinform. 2009;10(2):153–63.

26. Marshall MS, Boyce R, Deus HF, Zhao J, Willighagen EL, Samwald M, et al. Emerging practices for mapping and linking life sciences data using rdf - a case series. Web Semantics: Sci, Services Agents World Wide Web. 2012;14(0):2–13. Special Issue on Dealing with the Messiness of the Web of Data.

27. Luciano J, Andersson B, Batchelor C, Bodenreider O, Clark T, Denney C, et al. The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. J Biomed Semantics. 2011;2(Suppl 2):.

28. Abler D, Crichton C, Welch J, Davies J, Harris S. Models for forms. In: Proceedings of the Compilation of the Co-located Workshops on DSM'11, TMC'11, AGERE!'11, AOOPES'11, NEAT'11, & VMIL'11. NY, USA: ACM New York; 2011. p. 13–8.

29. Bioportal Ontologies http://bioportal.bioontology.org/.

30. Sim I, Carini S, Tu S, Wynden R, Pollock BH, Mollah SA, et al. The human studies database project: federating human studies design data using the ontology of clinical research. AMIA Summits Transl Sci Proc. 2010;2010:51–5.

31. Ogbuji C. A framework ontology for computer-based patient record systems. In: Proceedings of the ICBO: International Conference on Biomedical Ontology; 2011. p. 217–23.

32. Megan Kong Y, Dahlke C, Xiang Q, Qian Y, Karp D, Scheuermann RH. Toward an ontology-based framework for clinical research databases. J Biomed Informatics. 2011;44(1):48–58.

33. Samwald M, Jentzsch A, Bouton C, Stie Kallesøe C, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. J Cheminformatics. 2011;3(1):19.

34. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, et al. Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today. 2012;17(21):1188–98.

35. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J Biomed Informatics. 2008;41(5):706–16.

36. Pence HE, Williams A. Chemspider: An online chemical information resource. J Chem Educ. 2010;87(11):1123–4.

37. Willighagen EL, Alvarsson J, Andersson A, Eklund M, Lampa S, Lapinsh M, et al. Linking the resource description framework to cheminformatics and proteochemometrics. J Biomed Semantics. 2011;2(S-1):6.

38. Bilder G. Content negotiation for CrossRef DOIs. 2011. http://crosstech.crossref.org/2011/04/content_negotiation_for_crossr.html.

39. Laleci GB, Yuksel M, Dogac A. Providing semantic interoperability between clinical care and clinical research domains. Biomed Health Informatics, IEEE J. 2013;17(2):356–369.

40. Neumayr B, Schrefl M, Linner K. Semantic Cockpit: An Ontology-Driven, Interactive Business Intelligence Tool for Comparative Data Analysis In: De Troyer O, Bauzer Medeiros C, Billen R, Hallot P, Simitsis A, Van Mingroot H, editors. Advances in Conceptual Modeling. Recent Developments and New Directions SE - 9. Lecture Notes in Computer Science. Springer Berlin Heidelberg: Springer; 2011. p. 55–64.

41. Kamateri E, Kalampokis E, Tambouris E, Tarabanis K. The linked medical data access control framework. J Biomed Informatics. 2014;50(0):213–225.

42. PhUSE Wiki. http://www.phusewiki.org/wiki/index.php?title= Semantic_Technology.

43. FDA/PhUSE: Representing Existing CDISC Standards in RDF. https://github.com/phuse-org/rdf.cdisc.org.

44. NCI: CDISC Terminology. http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc.

45. Andersen M. Linked data to support clinical and non-clinical reporting. In: Proceedings of 2nd International Workshop on Semantic Statistics (SemStats 2014). Riva del Garda, Italy; 2014.

46. Williams T. A primer on converting analysis results data to RDF Data Cubes using free and open source tools. In: Proceedings of 10th Annual PhUSE conference(PhUSE 2014). London, UK; 2014.