Linked data to support Clinical and Non-Clinical Reporting

Marc Andersen

StatGroup APS, mja@statgroup.dk

Abstract. In pharmaceutical development Clinical and Non-Clinical studies are performed to develop new treatments and obtain regulatory approval for marketing. The current time and resource intensive reporting and regulatory review process may be facilitated by applying linked data principles. The RDF Data Cube vocabulary is evaluated for storing and documenting results, with the ultimate goal of complementing the current use of submitting results in PDF files by also providing the results and the associated metadata as linked data.

Keywords: Clinical data, results, metadata, CDISC, RDF Data Cube, statistics

1 Introduction

Clinical and Non-Clinical studies for development and registration of new treatments are performed, analysed and reported using standardised processes conforming to international and national ethical, legal and regulatory requirements.

Clinical trials are registered internationally [1] and assigned a unique identifier. In 2013 approximately 20.000 clinical trials were registered at U.S. ClinicalTrials.gov [2]. In general terms a clinical study is defined by a protocol and reported in a clinical study report [3]. Study data are represented using the Clinical Data Interchange Standards Consortium (CDISC [4]) standards, with increasingly standardised methods for analysis and reporting [5]. Results are delivered as Tables, Figures and Listings for inclusion in the clinical study report and submitted to regulatory authorities as PDF files [6]. While the study data are submitted in machine readable format, the trial results are currently not available in machine readable format. This offers a unique opportunity to complement the current approach with linked data methods.

This paper presents the experience of the Pharmaceutical Users Software Exchange (PhUSE) sub-team for Results Metadata [7] during the starting phase of developing a semantic representation of statistical results based on RDF and OWL. PhUSE is an independent, not-for-profit organisation run by volunteers with professional background such as Data Managers, Biostatisticians, Statistical Programmers and eClinical IT professionals.

2 Representation of Results

A typical clinical study report contains 100+ tables containing statistical analyses and descriptive statistics. A table contains from 10 to 1000+ results. In the context of this paper, a Result is defined as a data element derived by applying a statistic or statistical analysis to a set of source data.

The Results Metadata sub-team decided to start by limiting the scope to tables presenting descriptive statistics. The proposed approach is reflected in the bottom part of Figure 1, and incorporates derivation of results from source data.

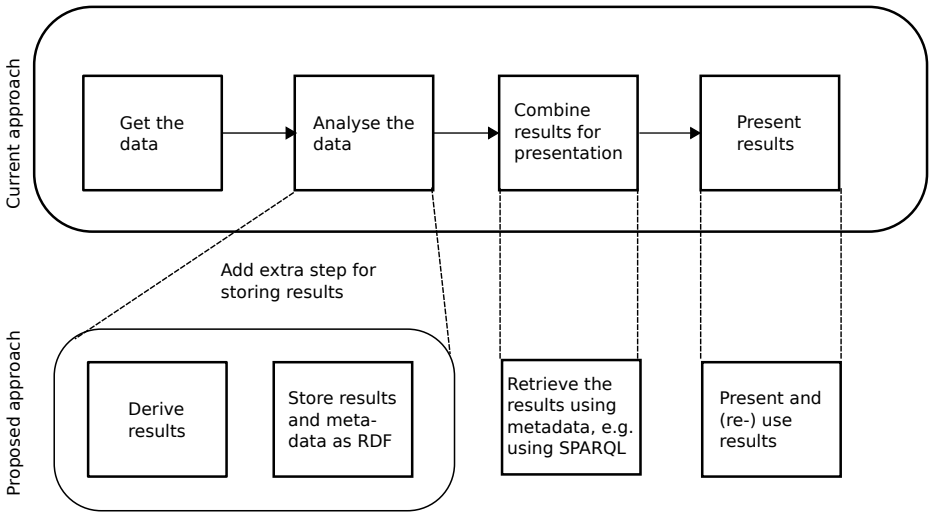

Fig. 1. Structure of a reporting program with proposed extension for linked data

2.1 RDF Data Cube for storing Results and associated Metadata

The main factors that influenced the sub-team's decision to investigate the RDF Data Cube [8] for representing Results are: 1) the vocabulary became a W3C Recommendation in January 2014; the Statistical Data and Metadata eXchange (SDMX [9]) has already been used for representing statistical data and 3) the RDF Data Cube has also been used for representing clinical data [10].

To get started with the RDF Data Cube the Publisci tool [11] and OpenRefine [12] and R [13] software tools were used. OpenRefine and R are discussed below. Ongoing tasks include determination of name spaces, determination of URI patterns [14, 15], derivation of code lists and linking the elements of the RDF Data Cube to the CDISC standards becoming available as RDF [16].

The sub-team decided to start with a basic approach of mapping each table in the clinical study report to an individual RDF Data Cube from results available

in a comma separated file (CSV). It was decided to use one RDF Data Cube observation for representing only one result (measure). This makes it possible to use the URI for referencing the results. Each statistics is identified using the cube dimensions and is described further with attributes. It was observed that the cube structure is not well suited in all circumstances: a table with descriptive statistics for weight by sex and for age overall, will have sex and parameter (age, weight) as dimensions, and will need a property for total over sex (like `sdmx-code:sex-T`). Another observation was that in addition to the existing consistency checks for a RDF Data Cube, specialised checks should be developed, such as verification of whether the value for total over sex is consistent with the values contributing to the total.

2.2 Producing RDF Data Cube from CSV data using OpenRefine

OpenRefine with the RDF plug-in [17] can be used to generate RDF Data Cubes based on existing results, such as results in CSV data. OpenRefine builds a cube skeleton around the data. Cube architects can save time by applying skeletons previously built around similar data. Observations are defined and attached to the skeleton as a final step before export as RDF Turtle or RDF/XML. OpenRefine is a good tool for defining and building cube structures. However, it is an unlikely candidate for the mass production of RDF Data Cubes, as it will be simpler to integrate the building of RDF Data Cubes in the usual reporting process.

2.3 Producing RDF Data Cube from source data using R

The RDF Data Cube can be generated in R with the R `rrdf` package [18], which uses Jena libraries [19]. This has the advantage of being able to derive results by extending programs with generation of RDF Data Cubes. For example, the RDF triple representing the value of the mean age for the subjects in the placebo group of the safety analysis population can be built from the `adsl` dataset as follows:

```
add.data.triple(qbstore,
  subject="ds:obs42",
  predicate="prop:measure",
  data=paste(
    mean( adsl[adsl$TRT01A=="Placebo" & adsl$SAFFL=="Y",]$AGE ) ),
  type="double" );
```

3 Conclusion

Based on the initial learning experience it appears practically feasible to enhance the current process of reporting and submission of clinical and non-clinical data following linked data principles. Providing clinical trial results as linked data will facilitate traceability, data sharing and integration, data mining and meta-analysis benefitting industry, regulatory authorities and the general public.

Acknowledgements

The author would like to thank all members of the Results Metadata team for input and discussion. The article represents the authors own interpretation. Special thanks go to Marcelina Hungria and Tim Williams for review and suggestions for improvements.

References

1. WHO International Clinical Trials Registry Platform (ICTRP) http://www.who.int/ictrp/en
2. U.S. ClinicalTrials.gov http://www.clinicaltrials.gov
3. ICH E3, Structure and Content of Clinical Study Reports, Finalised Guideline: November 1995 http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html
4. Clinical Data Interchange Standards Consortium (CDISC), http://www.cdisc.org
5. Li C, Bauer N. The Major Impacts of CDISC on Clinical Data Lifecycle. CDISC eJournal. http://www.cdisc.org/system/files/all/article/application/pdf/major_impacts_of_cdisc_on_clinical_data_lifecycle_li__bauer.pdf
6. Guidance for Industry Providing Regulatory Submissions in Electronic Format - General Considerations FDA, January 1999. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM072390.pdf
7. Results Metadata team, a sub-team of the PhUSE/FDA CSC Emerging Technology Working Group, http://www.phusewiki.org/wiki/index.php?title=Analysis_Results_Model
8. The RDF Data Cube Vocabulary. W3C Recommendation 16 January 2014 http://www.w3.org/TR/vocab-data-cube
9. Statistical Data and Metadata Exchange (SDMX) http://sdmx.org
10. Lefort L, Leroux H. Design and generation of Linked Clinical Data Cubes. 1st International Workshop on Semantic Statistics. 2013 https://www.ict.csiro.au/staff/laurent.lefort/SemStats2013-Lefort.pdf
11. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T. BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics. 15. October 2010;26(20):2617-9. http://dx.doi.org/10.1093/bioinformatics/btq475
12. OpenRefine http://openrefine.org
13. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
14. 223 Best Practices URI Construction http://www.w3.org/2011/gld/wiki/223_Best_Practices_URI_Construction
15. Capadisli S. Linked SDMX Data, 2013 http://csarven.ca/linked-sdmx-data
16. CDISC data standards in RDF. https://github.com/phuse-org/rdf.cdisc.org
17. RDF Refine - a Google Refine extension for exporting RDF. http://refine.deri.ie
18. Willighagen E. R package for handling RDF data, 2014. https://github.com/egonw/rrdf
19. Apache Jena. A free and open source Java framework for building Semantic Web and Linked Data applications. http://jena.apache.org