Paper TT01

Building a F.A.I.R. Foundation for Pharma: The PHUSE GoTWLD Project

Tim Williams, UCB Biosciences Inc., Raleigh, USA

ABSTRACT

Innovative pharmaceutical companies are rapidly adopting the FAIR guiding principles of Findable, Accessible, Interoperable, and Reusable data and standards to solve the many challenges facing our industry. Linked Data Knowledge Graphs provide the foundation for FAIR data, with potential impacts ranging from pre-clinical and late-phase studies to Real World Evidence, product launch, and sales. PHUSE initiatives under the umbrella of "Going Translational With Linked Data (GoTWLD)" employ FAIR principles to facilitate cooperative development in the pre-competitive space. Adoption of these principles will increase standardization and encourage data access and reuse for the benefit of companies, agencies, and patients.

WHAT IS FAIR?

At the heart of FAIR is Five Star Linked Open Data, a concept promoted by Sir Tim Berners-Lee as part of his original vision for the Semantic Web (1).

Table 1. Five Star Linked Open Data

★	**Available on the web, in any format, with an open license.**
★★	**Available as machine-readable structured data.** Example: Excel spreadsheet of data instead of a scanned image of a table of data.
★★★	**Non-proprietary format.** Example: A comma-separated value file instead of an Excel spreadsheet.
★★★★	**Open standards from the World Wide Web Consortium (W3C) to identify things.** Example: Identifiers using Resource Description Framework (RDF).
★★★★★	**Link your data to other people's data.**

The FAIR Guiding Principles (**Table *2***) (2) are receiving a lot of recent attention. Details are available in a host of other publications, including detailed descriptions of the concepts (3) (4).

Table 2. FAIR Principles

FINDABLE	F1. (Meta)data are assigned a globally unique and persistent identifier F2. Data are described with rich metadata (defined by R1 below) F3. Metadata clearly and explicitly include the identifier of the data they describe F4. (Meta)data are registered or indexed in a searchable resource
ACCESSIBLE	A1. (Meta)data are retrievable by their identifier using a standardized communications protocol A1.1 The protocol is open, free, and universally implementable A1.2 The protocol allows for an authentication and authorization procedure, where necessary A2. Metadata are accessible, even when the data are no longer available
INTEROPERABLE	I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (Meta)data use vocabularies that follow FAIR principles I3. (Meta)data include qualified references to other (meta)data

REUSABLE R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (Meta)data are released with a clear and accessible data usage license
R1.2. (Meta)data are associated with detailed provenance
R1.3. (Meta)data meet domain-relevant community standards

HOW FAIR IS PHARMA?

It is beyond the scope of this paper to detail how each FAIR component is built on the 5-Star concepts or to evaluate how well each principle is being adopted by the pharmaceutical industry. Rather, a single principle from each category is briefly discussed in this section.

FINDABILITY: F1. (META)DATA ARE ASSIGNED A GLOBALLY UNIQUE AND PERSISTENT IDENTIFIER

As our industry strives to obtain a more holistic picture of the patient to ensure successful treatment and follow up, the importance of unique and persistent identifiers, while maintaining data confidentiality, becomes increasingly important. Identifying the same subject across multiple trials using Unique Subject Identifiers (USUBJID) is a noble concept that often breaks down in real-world application due to data entry errors or transcription errors or a failure to recognize when the same patient participates in multiple trials. The use of a Linked Data Internationalized Resource Identifiers (IRIs), a type of Uniform Resource Identifier (URI) (5), would go a long way toward solving this problem.

ACCESSIBILITY: A2. METADATA ARE ACCESSIBLE, EVEN WHEN THE DATA ARE NO LONGER AVAILABLE

Traceability across the data lifecycle is a challenge that can be solved with existing technology. Currently, Biostatisticians and Medical Writers have difficulty accessing the metadata describing data collection and transformation processes. The nuances of these steps and the specifics of study design and protocol implementation can all be represented as machine-readable, human-interpretable data when Linked Data principles are applied.

INTEROPERABILITY: I1. (META)DATA USE A FORMAL, ACCESSIBLE, SHARED, AND BROADLY APPLICABLE LANGUAGE FOR KNOWLEDGE REPRESENTATION.

Clinical trials data and metadata lack a common language for representation. Historically, the industry has relied on a small number of vendors who provide proprietary solutions at a high cost. This is changing, albeit slowly, compared to other industries. A search in August 2019 on the job site Monster.com for "Biostatistician" yielded 1810 openings, 721 of which also listed R as a desired skill, and 230 for Python. Thinking beyond the traditional programming languages and applications we use to analyze and present our data, we need to consider how that data is represented *as knowledge*. Linked Data provides the semantics and integrated metadata for end-to-end data representation. As an open standard, Resource Description Framework (RDF) supports FAIR principles and can provide the shared knowledge representation when standards are openly published (in their entirety) and freely available to everyone.

REUSABILITY: R1.3. (META)DATA MEET DOMAIN-RELEVANT COMMUNITY STANDARDS

The move toward standardization has been extremely beneficial for our industry. Standards are available for pre-clinical and clinical trials from initiation of the study design through execution, analysis, and submission to regulatory authorities. Increasingly, these standards are facilitating data integration, including data from non-traditional sources like Real World Evidence (6). New use cases and the evolving nature of standards has revealed limitations in how standards are modeled. It is an easy task to take existing standards and convert them to Linked Data, but this approach would only compound the problems we face. To create a future-proof Knowledge Graph representation, a change in the approach is needed. The clinical trial process itself should be modeled and then standards applied to the types of entities, their relationships, and processes. This modelling should follow FAIR principles in an open, collaborative environment with results freely available to ensure development of robust models and subsequent adoption throughout the industry.

FAIRNESS SUMMATION

While there is much talk about FAIR in Pharma, there remain significant roadblocks to its adoption. As is the case for Semantic Web Linked Data, much development work can occur "behind the firewall" when applying open standards to proprietary and confidential data. Graph Databases are "going mainstream" (7), solving problems of metadata management, master data management, and supporting FAIR principles overall. Our risk-averse industry can catch up if companies, standards organizations, and regulatory agencies increase their participation in open, collaborative projects like those within PHUSE, a few of which are described in the next section.

BUILDING PHARMA'S FAIR FOUNDATION

Gartner states "The application of graph processing and graph DBMSs will grow at 100 percent annually through 2022…" (8). Are pharmaceutical companies prepared to grow their skill set accordingly, to leverage Linked Data technology projects based on FAIR principles? Presentations at the Semantics@Roche conference in April 2019 identified multiple initiatives. Bayer presented their novel approach for engaging internal staff in FAIR principles (9) and in his keynote address, Dr. Mark Musen highlighted the CEDAR project for metadata (10). Addressing the lack of available tooling, in late 2019 the Pistoia Alliance will launch a freely available toolkit (11) to help companies adopt the FAIR guiding principles. All signs point toward a growing commitment to FAIR.

The pharmaceutical industry continues to move toward more open collaboration in the pre-competitive space, with the new PHUSE project "Open Source Technologies in Clinical Research" (12) as a recent example. The Roofshot/Moonshot Manifesto proposed at EUConnect18 (13) for adoption of Linked Data was embraced by the PHUSE project "Clinical Trials Data as RDF" and that project was reborn as "Going Translational with Linked Data (GoTWLD)." GoTWLD consists of three Roofshot initiatives described in the next section. Roofshots are smaller, incremental steps which build toward Moonshots. Moonshots strive to apply and invent state-of-the-art solutions.

ROOFSHOT 1: MEDDRA AS RDF

GitHub Repository: https://github.com/phuse-org/MedDRAasRDF

The Medical Dictionary for Regulatory Activities (MedDRA) is clinically validated and is used primarily by the International Conference on Harmonization (ICH) and its members to facilitate standardized information sharing. The GoTWLD project converted MedDRA terminology to RDF to characterize adverse events in the project's test case data conversions.

Companies have independently converted MedDRA to RDF. There is no published, shared methodology for these efforts, leaving each company to re-invent the process. The MedDRAasRDF project provides a data model for MedDRA terminology and scripts for converting the multiple ASCii files to RDF. MedDRA is licensed, so source files are not provided in the GitHub repository. The project was granted a Research/Non-Commercial license from the MSSO in order to perform this work. The conversion scripts can be freely used by anyone to construct the RDF from licensed copy of terminology files.

The Linked Data approach provides a single terminology file that maintains the terminology hierarchy using meaningful, explicit, semantic relationships. The graph data model makes it easy to include additional metadata and documentation. A future improvement would be the inclusion of language tags, which would make translation seamless and virtually instantaneous.

Although the terminology files are licensed and are therefore not FAIR, the data model and conversion process are freely available for download from the GitHub site. The team encourages industry, regulatory, and standards organizations to contribute to the project. The MSSO has expressed interest in the project and we hope that future terminology releases may be available as RDF or via a secure SPARQL endpoint API. In September 2019 the project was picked up by ROpenSci as an example use case[1] for the R package rdflib.

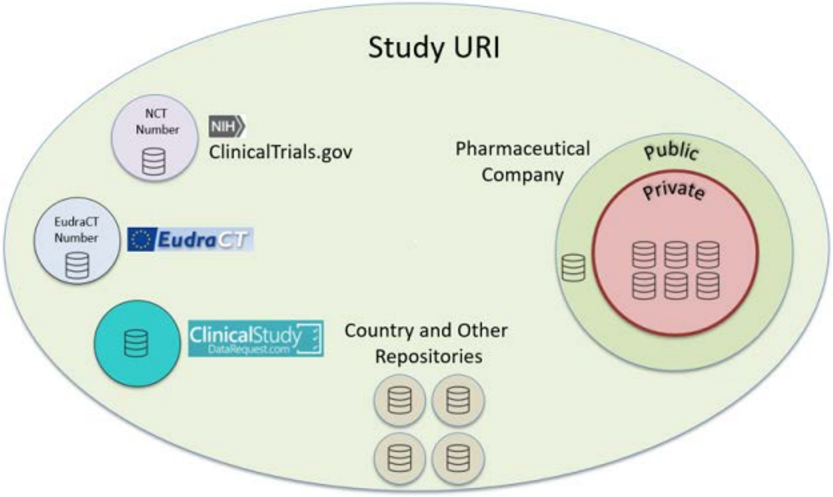
ROOFSHOT 2: UNIQUE IDENTIFIERS FOR THE PHARMACEUTICAL INDUSTRY

GitHub Repository: https://github.com/phuse-org/UIDPharma

"Unique Identifiers for the Pharmaceutical Industry" is based on the PhUSE EUConnect18 paper Study URI (14) and is presented as a poster at EUConnect19. While the concept of unique identifiers for clinical trials is not new, implementations are inconsistent and subject to change over time. The NCT Number from ClinicalTrials.gov (15) is one of the most widely recognized study identifiers; however, there are limitations. One limitation is the fact that NCT numbers are created after submission and review of the protocol at ClinicalTrials.gov, so it is not available earlier in the

[1] https://twitter.com/rOpenSci/status/1171800566948929537

development process. The European Medicines Agency provides another unique identifier for trials, called the EudraCT number, which can be used to look up information in the European Clinical Trials Database (EudraCT). There are also repositories for specific countries, as well as multiple databases within the pharmaceutical companies themselves. There is currently no standard way to bring this information together for patients, researchers, or regulatory organizations.



The GitHub repository and EUConnect19 poster provide information on proposed structure, recommended predicates, and technical considerations for how to create and use URI's with concepts that go well beyond study identification. Global unique and persistent identifiers will go a long way toward facilitating data integration and providing timely information to researchers and patients searching for trials.

ROOFSHOT 3: A LINKED DATA APPROACH TO SEND CONFORMANCE (SENDCONFORM)

GitHub Repository: https://phuse-org.github.io/SENDConform/

SENDConform is a proof of concept for representing SEND conformance checks using Shapes Constraint Language (SHACL)[2]. The goal is to increase efficiency by decreasing labor-intensive checks performed by both study sponsors and regulatory agencies. The project builds upon the ontology and data conversion work in the SDTM domain that was part of the original "Clinical Trials Data as RDF" project and extends that work into pre-clinical data. SHACL constraints developed for pre-clinical data may later support SDTM data checks.

The goal of the project is to model FDA SEND Validation Rules for the demographics (DM) and Trial Summary (TS) domains. By constructing constraints in a modular fashion and using a data schema similar to SDTM, rules in RDF can be efficiently created, maintained, and deployed. FDA rules are deconstructed into their primary components and modeled in SHACL with support for RDF inferencing. These rules can form the basis of data capture applications that warn of potential data problems during entry and can be used to automate data quality assessments at the pharmaceutical company or regulatory agency. Working collaboratively in the pre-competitive space to develop these checks ensures they follow FAIR principles.

CONCLUSION

FAIR can play a significant role in the pharmaceutical industry's future, but only after significant obstacles are overcome. These obstacles exist from the highest level of the organization all the way down to the individual. When standards and data checks become open, companies and individual programmers are freed to explore new opportunities based on providing high quality, integrated data to consumers. Data capture, including metadata creation and management, can become increasingly automated as we move toward end-to-end data capture and management processes.

[2] https://www.w3.org/TR/shacl/

References

1. **Berners-Lee, Tim.** Linked Data. *World Wide Web Consortium.* [Online] 07 27, 2006. [Cited: 08 19, 2019.] https://www.w3.org/DesignIssues/LinkedData.html.
2. FAIR Principles. *Go-FAIR.* [Online] [Cited: 08 19, 2019.] https://www.go-fair.org/fair-principles/.
3. al., Mark D. Wilkinson et. The FAIR Guiding Principles for scientific data management and stewardship. *Nature.* [Online] 03 15, 2016. [Cited: 08 19, 2019.] https://www.nature.com/articles/sdata201618.
4. FAIR data. *Wikipedia.* [Online] [Cited: 08 19, 2019.] https://en.wikipedia.org/wiki/FAIR_data.
5. Uniform Resource Identifier. *Wikipedia.* [Online] [Cited: 08 19, 2091.] https://en.wikipedia.org/wiki/Uniform_Resource_Identifier.
6. [Online] US Food and Drug Administration, 05 09, 2019. [Cited: 08 19, 2019.] https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.
7. Cagle, Kurt. Graph Databases Go Mainstream. *Forbes.com.* [Online] 07 18, 2019. [Cited: 08 19, 2019.] ? https://www.forbes.com/sites/cognitiveworld/2019/07/18/graph-databases-go-mainstream/#4a2d2a8a179d.
8. Gartner, Inc. Gartner Identifies Top 10 Data Analytics Technology Trends for 2019. . [Online] [Cited: 08 19, 2019.]
9. *Making Data FAIR at Bayer.* al., Alexandra Grebe de Barron et. Basel : Semantics@Roche, 2019.
10. CEDAR : Better metadata means better science. [Online] Stanford University. [Cited: 078 19, 2019.] https://metadatacenter.org.
11. The Pistoia Alliance to Develop a Toolkit to support Implmentation of the FAIR Guiding Principles. *Pistoia Alliance.* [Online] 07 17, 2019. [Cited: 08 19, 2019.] https://www.pistoiaalliance.org/news/fair-toolkit/.
12. Open Source Technologies in Clinical Research. [Online] PHUSE. [Cited: 08 19, 2019.] https://www.phusewiki.org/wiki/index.php?title=Open_Source_Technologies_in_Clinical_Research.
13. *Overcoming Resistance to Technology Change: A Linked Data Perspective.* Williams, Tim. Franfurt : PHUSE, 2018. TT04.
14. *Study URI.* Kerstin Forsberg, Daniel Goude. Frankfurt : PHUSE, 2018. TT09.
15. ClinicalTrials.gov. [Online] [Cited: 08 19, 2018.] https://clinicaltrials.gov/.

ACKNOWLEDGMENTS

The author is indebted to the contributions from the "GoTWLD" PhUSE project team members, especially Dr. Armando Oliva for his ontology work.

This paper is largely based on free, open-source software and the efforts of volunteers in PhUSE working groups. Please support those who donate their time and expertise through your own collaboration, participation, and promotion of these activities.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Tim Williams
UCB BioSciences, Inc
Raleigh, NC, USA
tim.williams@ucb.com
@NovasTaylor
https://www.linkedin.com/in/timpwilliams

Brand and product names are trademarks of their respective companies.