

Overcoming Resistance to Technology Change: A Linked Data Perspective

Tim Williams, UCB Biosciences Inc., Raleigh, USA

ABSTRACT

Deployment of technology in a new arena encounters many obstacles. Companies frequently retain proven but antiquated tools to reduce risk. A conservative mindset is understandable in the pharmaceutical industry, where patient safety is paramount, regulatory submission guidelines must be followed, and product failures are costly. However, if an industry becomes too comfortable with the status quo, it may fail to identify the risks associated with *not* modernizing, instead choosing to focus on supporting obsolete systems and processes.

Linked Data as Resource Description Framework (RDF) faces numerous implementation challenges despite its immense potential for our industry. This paper explores the issues around its adoption and provides strategies that can also be applied to other technologies being introduced into our industry for the first time.

INTRODUCTION

The gap between the technology evangelists and those responsible for approvals and implementation is particularly noticeable for Linked Data in the pharmaceutical industry. My original title for this paper was "Bridging the Gap Between Linked Data Nerds and Everyone Else." I observed that Linked Data presentations at recent PhUSE conferences are attended by primarily two groups of people: Linked Data Nerds (LDNs) who are technology-savvy and a small group of the curious. Failure to reach beyond a specialist audience slows uptake of any technology entering our industry for the first time, so I took my own advice and changed the paper's title.

Semantic Web¹ technology, built on the concept of Linked Data² as Resource Description Framework (RDF³), is not new, but its implementation in our industry is recent. Concepts and terminology in this field is confusing, often misused, and regularly debated by experts. For the purpose of this discussion, the term *graph database* includes both *RDF graphs* and *labeled property graphs* like Neo4j, as recently articulated by A.I. Hatzis:

"A Graph Database is a database that uses a graph topology, i.e. vertices and edges, to manage information at the conceptual level independent of the logical and physical implementation of the graph data structure." - Athanassios I. Hatzis, 28th February 2017⁴

I reserve the contentious term *knowledge graph*⁵ for RDF graph databases that use ontologies to organize the entities, relationships, and rules in the data. Ontologies facilitate inferencing of additional information using a semantic reasoner⁶. *Knowledge graph* has an intuitive appeal the wider non-technical audience and should not be confused with the Google Knowledge Graph⁷ initiative. I use *Linked Data* to refer to data as RDF, with or without the use of ontologies.

This paper does not provide a roadmap for transitioning a company to Linked Data. Change management requires knowledge of the current baseline and a well-defined future state. Linked Data faces many unknowns in our industry. Will regulatory agencies accept it as a valid submission format? Will Electronic Data Capture (EDC) systems capture data natively as RDF or will semantic technologies function solely as a bridge between data silos? Fortunately, it is not necessary to have all the answers at the outset. Graph database models are flexible and can be developed and extended over time.

PhUSE EU Connect 2018

Comments at three PhUSE conferences led to the conception of this paper. Each began with the same (paraphrased) statement that is addressed in corresponding sections of the paper:

"Linked Data is really cool, but..."

"...I can already do that in a relational database." - USConnect 2018

A. TECHNOLOGY

"...I don't see how to apply it to what we do." - US CSS 2018

B. APPLICATION

"...it will never happen in Pharma." - EUConnect 2017

C. ADOPTION

A. TECHNOLOGY

Can relational databases provide the same functionality and superior performance compared to graph databases? Proponents of opposing relational and graph systems question the other's knowledge and the discussion deadlocks. A full comparison of the two is beyond both the scope of this paper and my personal expertise.

While traditional clinical trial results are neither high-velocity nor transactional in nature (both strengths of relational databases), the vast number of required nodes and relations raises storage and performance concerns. The continuing trend of decreasing storage costs, along with the ability to easily expand and dynamically allocate storage, makes this less of an issue than it was just a few years ago.

Graph databases are now highly performant. In contrast to graph databases, complex joins across multiple tables can greatly decrease performance in relational systems. Globally unique identifiers in graph databases remove the use of joins and foreign keys. One version of the truth is stored one time, in one place. It is not necessary to define the entire schema in advance and it is much easier to change and extend it later. The database schema, metadata, and instance data are all stored together as triples and queried using SPARQL⁸. The graph data model is intuitive, with data and processes modelled the way they exist in the real world. There is no need for a separate metadata repository and it is possible to integrate vastly diverse types of data⁹.

The use of Web Ontology Language (OWL¹⁰) is a distinct advantage for RDF¹¹, enabling interpretation of data and its meaning by both human and machine. OWL supports ontology development and reasoning that identifies data problems and infers additional information based on existing data and relationships. OWL defines the nature of the relationship between two entities or between an entity and a literal value (string, integer, et cetera). By bestowing interpretable meaning on the entities in the model, ontologies provide more than classification, structure, and rules.

Additional comparisons of relational with graph databases are found online, with biases from each side of the discussion. Companies will not replace existing relational systems and skill sets because of the expense. Both technologies will coexist, leveraging the strengths of each. Graph databases will provide knowledge management and platforms for merging disparate sources of data from traditional systems, gradually expanding their role behind the scenes, while interfaces and data models are developed. We need to move past the adversarial *relational versus graph databases* arguments and instead explore how each one compliments the other in a comprehensive data management strategy. Your all-star data management team should now include Ontologists and Enterprise Knowledge Engineers. (1)

B. APPLICATION

Failure of the LDNs to effectively communicate how Linked Data can provide practical solutions contributes to the perception that "it will never happen" (**section C**). Their technical focus must widen to include the business perspective and differing levels of expertise.

"It almost seemed as if the authors intentionally wanted to encourage the perception of high priesthood, only to be fathomed by the chosen few." - David McComb¹²

Elitism within the Linked Data community has hindered broader acceptance. Individuals prototyping practical solutions have been criticized for using buzz words and not fully understanding the technology landscape. When understanding is lacking, a more constructive approach is to educate and facilitate practical innovation with positive encouragement, dispelling the myth that the field is too complex for non-academics.

PhUSE EU Connect 2018

"People think RDF is a pain because it is complicated. The truth is even worse. RDF is painfully simplistic, but it allows you to work with real-world data and problems that are horribly complicated." - Dan Brickley and Libby Miller (2)

To appreciate the potential impact of Linked Data requires a change in mindset. When you hear the word "data", you likely picture a row and column structure like an Excel table, SAS dataset, R data frame, or a relational table in a database. This has been our collective experience as data managers, programmers, statisticians, data scientists, et cetera. Our thinking needs to shift to one of an interconnected graph of data points and the relationships between them. Transforming two-dimensional, row and column thinking to a networked graph of data takes time. LDNs too often model real-world entities and relationships without fully exploring the impact on real-world processes. Examples must be customized to each audience (to facilitate acceptance and understanding) and answer their most important question: "How does this solve *my* problem?"

The fondness LDNs have for illustrating graph data models using large interactive force network graphs often has the opposite of the desired effect. The uninitiated may see these intricate visuals as hopelessly convoluted and complex in contrast to familiar row and column structures. It offers little comfort when LDNs suggest hiding this complexity behind user-friendly interfaces that offer superior ways of working with the data. The architects of the labeled property graph Neo4j¹³ addressed usability early in their product development. They built a user-friendly interface that lowered the bar of entry for inexperienced users. Vendors of RDF databases must learn from this example. As a case in point, the PhUSE Linked Data Interactive Workshop (3) needed an easy-to-use whiteboard interface for RDF (similar to Neo4j) and discovered it was necessary to build it from the ground up.

A key factor for success is resisting the temptation to translate current standards and data 'as-is' to RDF and instead opt to invest in developing ontologies to represent the entities and relationships present in the clinical trial process. Data models backed by ontologies are flexible, extensible, and not reliant on specific versions of standards. The industry currently creates datasets that comply to a specified version of a standard, then spends vast amounts of time and money converting between versions. As the size and the diversity of data grows, the row and column model becomes untenable. Standards like SDTM and ADaM should be applied as views or templates that materialize the data into use-case dependent structures. In this future scenario, both the standards and the data are stored as RDF. In the short-term, creation of Define-XML becomes virtually automatic (and eventually unnecessary)¹⁴ when ontologies, rules, metadata, and data are all stored together. In the longer-term, SDTM and ADaM (as we know them) will cease to exist, replaced by superior, on-demand, fit-for-purpose data constructed from a graph database. There will be no need for a separate Define document because definitions, metadata, meaning, and documentation will be integral to the data.

An excellent example of the effectiveness of RDF for managing SDTM terminology and its impact on the modeling and use of biomedical concepts was presented by David Ibersen-Hurst at PhUSE USConnect18. (4) The original specification of Blood Urea Nitrogen (BUN) was replaced in later versions of SDTM terminology by UREAN, which can specify if a serum (C13325), plasma (C13356), or serum or plasma (C105706) specimen was used for the analysis. The change affects both the terminology used in SDTM datasets and how the concept is modeled. Linked Data provides an effortless way to both detect the change and map data between the versions (BUN in older data maps to UREAN in the new data). If RDF had originally been used to *model how the values are collected in real-world scenarios* (serum, plasma, serum or plasma), it is likely that the UREAN approach would have been used from the outset. Even if it were not, changing a process to accommodate both new and old versions of the terminology is much easier in a graph database than in a relational database.

Converting from row and column data to RDF results in higher quality data. Ambiguous, problematic, or missing data has nowhere to hide thanks to Linked Data's unique identifiers and rules. Converting to graph data reveals problems you were not previously aware of and confidence in your data quality may drop. Teams must keep in mind the benefits of the conversion process when going through the steps needed to improve data quality, recognizing that the problems existed in the data the entire time and are only now being identified.

THE RETURN ON INVESTMENT (ROI) UNICORN

Only a few pharmaceutical companies have prepared business cases for Linked Data, which may in part be due to the disruption it would cause to existing processes. The wide-ranging impact of the introduction of the shipping container on the freight industry (5) provides a thought-provoking parallel. When data is viewed as a commodity that is created, formatted, packaged, transported,

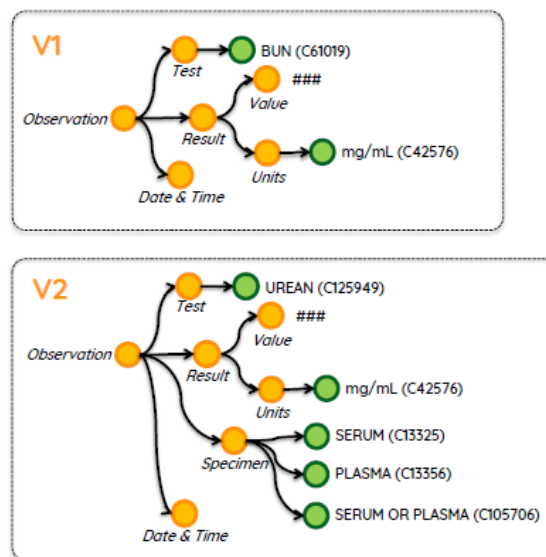


Figure 1 SDTM Terminology for Urea Nitrogen
Image used with permission from David Ibersen-Hurst, A3 Informatics. (3)

PhUSE EU Connect 2018

and then used for various purposes, we realize how Linked Data impacts every aspect of the data lifecycle. The far-reaching and disruptive effect of these changes makes it difficult to calculate ROI. The optimistic message of LDNs that Linked Data will positively impact the entire data lifecycle needs to be moderated when communicating to managers, who expect fast solutions with known ROI in a validated environment. It is nearly impossible to delineate a business case for the entire lifecycle, precisely because the change is so disruptive.

Consider the role of RDF's use of unique identifiers¹⁵ in ROI estimates. Unique identifiers help to ensure that an entity is created once and then referenced throughout the entire lifecycle, ensuring accuracy and eliminating costly and error prone duplication. For example, a treatment intervention defined in the *study design* is also needed in the trial *protocol*, has implications for *data capture* (*Case Report Forms*, data capture systems), is a part of the *results data*, and appears yet again *submissions* and *publications*. The obvious advantage of creating a single data value that is referenced multiple times exposes the complexity of calculating ROI across the entire lifecycle for that data point. One approach to this challenge is to calculate ROI around a well-defined problem that is isolated to a specific portion of the data lifecycle. A risk in this constrained approach is that the resulting ontologies and terminology must be later reworked for wider adoption; thus, continually "re-inventing wheel."

Data quality is a prime target for ROI. As reported at the PhUSE USConnect18 conference, 32% of submissions violate at least one conformance criteria (6) and 20% of upload attempts into the FDA Janus database fail¹⁶. A missing TS domain file or no study start date are the most common errors.⁽³⁾ When multiple studies are submitted, it is unclear to which study the TS file belongs. Date format errors are common, as are submission of Define-PDF in place of Define-XML, and confusion as to which .XPT file is associated with a Define-XML. These type of errors are removed when conformance requirements are coded within the graph data. However, this seemingly easy RIO calculation is complicated when companies do not accurately capture baseline numbers as basic as the number of hours spent on specific tasks.

Further complicating the calculation is the fact that adopting new technology¹⁷ creates a gap between legacy systems and the future state. The significant costs of changing IT platforms, processes, and staff skill sets all decrease ROI in the short term, before the benefits are realized. Human factors lag behind the technological change, hampering adoption and further complicating ROI. Change is often viewed as a threat and must be handled deftly. As the builders of ontologies and converters of data, LDNs must work closely with business experts and systems architects. The pharmaceutical industry has a vast talent pool that is often walled-off both within and between companies. It is important to identify and facilitate communication between internal and external visionaries. Because domain knowledge is integral to the data, you must reach far beyond the usual technology expertise to include experts in both clinical and non-clinical data.

When a compelling ROI calculation leads to project approval, the quest for a suitable database vendor begins. This is yet another challenging task due to the nature of Linked Data.

Re: semantics " you can't just buy a vendor product and catch up. Takes a long term investment in skills development." Dave Newman #EDW18 - Dave McComb
(@semanticarts) Twitter, 25 April 2018

Adopting Linked Data is a transformative data journey. Any one of a variety of triplestores¹⁸ may fit the purpose. A key success factor is development and use of the appropriate ontologies. Be wary of vendors with limited experience in the pharmaceutical domain who do not take the time to understand the standards, regulatory requirements, and other complexities surrounding pharmaceutical data.

It is most beneficial when Linked Data is employed at the start of the data life cycle. It also functions very well as an integration layer between data silos. However, its potential to "save the day" by integrating legacy data is jeopardized by the quality of the source data. Failure to address quality and completeness of source data merely transforms a relational problem into a graph problem. Data augmentation and modeling is also needed, making the conversion process much more than a simple upload to a new system.

PhUSE EU Connect 2018

DEVELOPMENT PHILOSOPHY

We find ourselves dreaming of the moon while designing the rocket that will take us there. Our predicament parallels Google's "moonshot factory," as succinctly described by Luiz André Barroso (7) and in **Figure 2**. The moon is enticing, where all data is interconnected across the data lifecycle, underpinned by comprehensive, machine-interpretable ontologies, rule sets, and standards. Although attractive, the attempted leap from the ground (relational database) to the moon (graph database) is devastatingly disruptive and impossible to achieve.

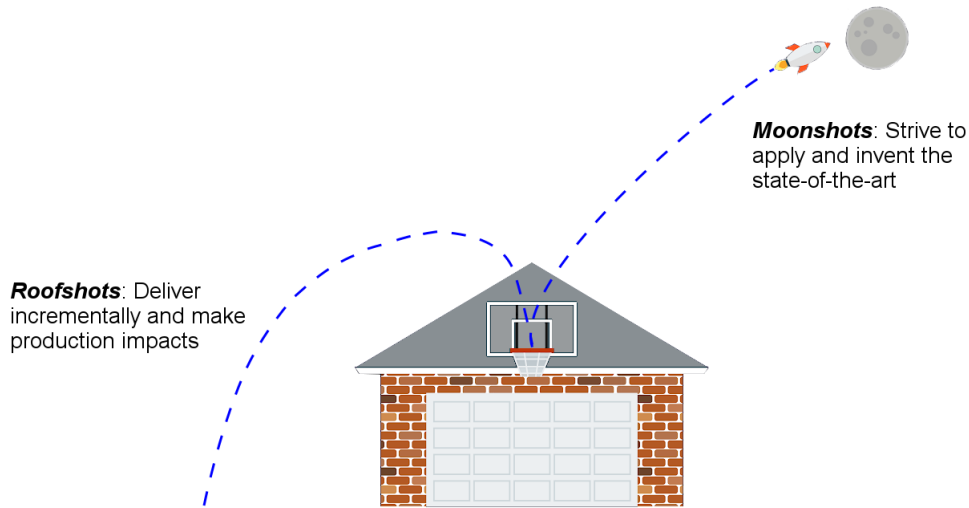


Figure 2 The Roofshot Manifesto
Original image from (7) with annotations from (8)

We need attainable, incremental roofshots while keeping our moonshot in sight. The challenges of managing clinical trials data can be broken down into projects with manageable scope and coordinating across teams as we develop shared ontologies. As an example, the Clinical Trials Data as RDF project started with a small number of SDTM domains (DM, VS, EX) and modeled the entities needed to represent the data *in just those domains* (a minor roofshot). Entities like treatment arm should be represented in a protocol ontology. Rather than attempting to develop a comprehensive protocol ontology for the project, the team chose to represent only those protocol entities needed in their current subset of SDTM domains. As more SDTM domains are added, the protocol ontology is amended with additional entities. Additional roofshot projects are anticipated in the clinical and non-clinical space. PhUSE projects must also leverage work from groups like CDISC (SDTM terminology as RDF as an example) and OBO Foundry¹⁹ (ontologies). As if by magic, the user experience must become one where data is easily accessed with provenance, metadata, and standards available. A moonshot to be sure, but achievable using many coordinated roofshots.

C. ADOPTION

The sentiment behind the statement that Linked Data "will never happen in the pharmaceutical industry" comes in part from a lack of identifying where and how to apply the technology (**section B**). Moreover, pessimism and a reluctance to change from legacy processes and applications is common place.

A prime example is the SAS V5 XPT transport file developed in the late 1980's²⁰ and identified by the FDA as the mechanism for data delivery for submissions in 1999. (9) This thirty-year-old file format is often quoted by both pharmaceutical companies and vendors as a reason for not updating systems and processes. It is understandable how vendors, with a vested interest in SAS transport files, propose a move from version five (V5) to version eight (V8) of the transport format as "modernization." Their proposal fails to address the critical shortcomings of the two-dimensional file format and would continue to band-aid legacy systems and data models that are in the twilight of their lifespan. Instead of looking to the past for answers and using regulatory requirements as an excuse to avoid change, the industry needs to look toward the future and outward to other industries for solutions. It is a myth that regulatory agencies are not interested in Linked Data and alternative submission formats. Even as agencies strive to maintain vendor neutrality and are rightfully reluctant to recommend technology, they too are eager to modernize. Once a critical mass of support develops around an XPT alternative, I am certain that both the regulatory agencies and the industry will jump at the opportunity.

PhUSE EU Connect 2018

Linked Data can produce better quality, submission-ready data and do so while supporting existing requirements and formats. As observed by Marshall McLuhan:

"...the old medium is always the content of the new medium." - Marshall McLuhan

The PhUSE project "Clinical Trial Data as RDF"²¹ is prototyping the creation of high-quality SAS V5 transport files and Define-XML from Linked Data, providing a bridge between the antiquated XPT format and the modern, multidimensional graph. Graph databases are also an excellent choice to provide an integration layer for legacy systems, using approaches like virtual graphs and R2RML²².

Collaboration within the industry becomes key, because successful Linked Data strategies will be based on ontologies. Companies must cooperate in ontology development, much like how the Study Data Tabulation Model (SDTM) standard is developed and promoted by CDISC²³. Standards organizations and regulatory agencies are unlikely to lead adoption of Linked Data, yet both are needed as stakeholders. Because the standards will be intimately intertwined with the results data, it is not simply a matter of converting the standards as they now exist to RDF. Once again, the entities within the standards must be modeled to represent real-world processes, entities, and rules. If an ontology-based approach is not adopted for the conversion of standards, the chance of success is low.

Organizations that foster cooperation between pharmaceutical companies in the pre-competitive space (e.g. PhUSE²⁴, TransCelerate²⁵) can act as sponsors and incubators. Progress is slow and sporadic in these organizations where volunteer attendance competes with other work and personal commitments. This is an especially difficult challenge for complex Linked Data initiatives that require a dedicated team to solve complex problems. Progress on Linked Data initiatives may require formation of a new group outside of existing organizations, with members from the broader Semantic Web community working with subject matter experts from the pharmaceutical industry. Much progress could be made if these experts were to come together in a dedicated symposium/hackathon focused on ontology development and implementation.

Linked Data has not received the same level of attention as other popular technologies, like Machine Learning, Artificial Intelligence, or Blockchain.

"I think graph databases have gotten the exact opposite of hype. A few vendors have more or less quietly delivered real value with techniques that [are] difficult to impossible with other technologies." - Neil Raden (@NeilRaden) Twitter 11April 2018

The lack of hype can benefit adoption by moderating expectations and providing time for ontology and user interface development. That time is now. Several pharmaceutical companies are quietly adopting graph databases, vendors are recognizing the value of metadata repositories based on RDF²⁶, and standards are increasingly available in RDF²⁷. The best counter point for "it will never happen" is the fact that it is, indeed, happening.

CONCLUSION

Should your company embark on the Linked Data journey? This question is best answered on a case-by-case basis. The industry's reliance on relational database systems must be augmented with graph solutions. Relational databases cannot provide the same functionality as knowledge graphs, but the latter is not a replacement for the former. A comprehensive data management strategy includes both. Plotting a transition must account for legacy systems, vendors, and processes. Disruption to ongoing operations must be minimized.

Companies that fully embrace the capabilities provided by Linked Data will achieve a profound competitive advantage. There will be significant failures along the way, due to choosing a wrong vendor, failing to define appropriate project scope, and a host of other reasons. Companies with a "it will never happen" technology philosophy are destined to fall behind their competitors. Pharmaceutical companies too often take the approach of "you first, I will follow", resulting in a lack of leadership. We must stop using regulatory agencies and vendors as excuses for inaction and aversion to innovation.

The sheer number of problems that can be solved by Linked Data paradoxically contributes to the reluctance to pursue it as a solution. All of the advantages of a Linked Data Knowledge Graph will be realized when an ontological approach is applied across the complete data lifecycle, which is both highly advantageous and highly disruptive. Linked Data *is* happening in our industry. The question is whether it can overcome the resistance to change that would prevent it from reaching its full potential.

PhUSE EU Connect 2018

REFERENCES

1. **Cagle, Kurt.** Why Most Companies Need An Ontologist (or Two). *LinkedIn*. [Online] 03 03, 2016. [Cited: 08 13, 2018.] <https://www.linkedin.com/pulse/why-most-companies-need-ontologist-two-kurt-cagle/>.
2. *Validating RDF Data.* **Gayo, Jose Emilio Labra, et al.** s.l. : Morgan & Claypool, 2018.
3. *Linked Data for Clinical Trials: An Interactive Hands-on Workshop.* **Tim Williams, Johannes Ulander.** Frankfurt : PhUSE, 2018.
4. *It's Time to Change.* **Iberson-Hurst, David.** Raleigh, NC : PhUSE USConnect, 2018.
5. **Levinson, Marc.** *The Box. How the Shipping Container Made the Whole World Smaller and the World Economy Bigger.* 2nd. s.l. : Princeton University Press, 2016.
6. *FDA View: Technical Rejection Criteria for Study Data.* **Chen, Ethan, et al.** Raleigh, NC : PhUSE, 2018.
7. **BARROSO, LUIZ ANDRÉ.** The Roofshot Manifesto. [Online] July 13, 2016. [Cited: July 13, 2018.] <https://rework.withgoogle.com/blog/the-roofshot-manifesto/>.
8. *Challenges and Innovations in Building a Product Knowledge Graph.* **Dong, Xin Luna.** Applied Data Science. Invited Talks at SigKDD : s.n., 2018.
9. *The Past, Present, and Future of Clinical Data Standards.* **Decker, Chris.** s.l. : SAS Global Forum, 2010. Paper 183-2010.
10. **Allemang, Dean and Hendler, Jim.** *Semantic Web for the Working Ontologist, 2nd Edition.* 2011.

ACKNOWLEDGEMENTS


This paper is largely based on the efforts of volunteers in the PhUSE organization and 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.

¹ https://en.wikipedia.org/wiki/Semantic_Web

² https://en.wikipedia.org/wiki/Linked_data

³ https://en.wikipedia.org/wiki/Resource_Description_Framework

⁴ <https://www.datasciencecentral.com/profiles/blogs/a-quick-guide-on-how-to-prevail-in-the-graph-database-arena>

⁵ [https://en.wikipedia.org/wiki/Ontology_\(information_science\)](https://en.wikipedia.org/wiki/Ontology_(information_science))

⁶ https://en.wikipedia.org/wiki/Semantic_reasoner

⁷ https://en.wikipedia.org/wiki/Knowledge_Graph

⁸ <https://www.w3.org/TR/rdf-sparql-query/>

⁹ <https://www.w3.org/2007/03/VLDB/>

¹⁰ https://en.wikipedia.org/wiki/Web_Ontology_Language

¹¹ Some labeled property graphs like Neo4j and other hybrid graphs can import RDF and employ reasoners.

¹² Forward to "Demystifying OWL for the Enterprise", commenting on the perceived complexity of the Semantic Web as a factor in its slow uptake.

¹³ <https://neo4j.com/>

¹⁴ Additional documentation like Define-XML becomes redundant when semantics (meaning) is integral to the data itself.

¹⁵ https://en.wikipedia.org/wiki/Internationalized_Resource_Identifier

¹⁶ Comment from audience member with knowledge of the upload process, attending paper RG09 "FDA View: Technical Rejection Criteria for Study Data", PhUSE USConnect18.

¹⁷ For the purpose of this paper refer to Linked Data as a "new technology" because adoption in the pharmaceutical industry is a relatively new event, even though the technology itself has been around for many years.

¹⁸ <https://www.w3.org/wiki/LargeTripleStores>

¹⁹ <http://www.obofoundry.org/>

²⁰ <https://www.loc.gov/preservation/digital/formats/fdd/fdd000464.shtml>

²¹ <https://github.com/phuse-org/CTDasRDF>

PhUSE EU Connect 2018

- ²² <https://www.w3.org/TR/r2rml/>
- ²³ <https://www.cdisc.org/standards/foundational/sdtm>
- ²⁴ <https://www.phuse.eu/>
- ²⁵ <http://www.transceleratebiopharmainc.com/>
- ²⁶ <https://www.a3informatics.com/> , <https://nurocor.com/>
- ²⁷ <https://www.cdisc.org/standards/data-exchange/rdf>