Semantic models for pharma

Connecting data across drug discovery

Ilaria Maresi [(1)], Jochem Bijlard [(1)], Kees van Bochove [(1)]

1. **The Hyve**, Arthur van Schendelstraat 650, 3511 MJ Utrecht, The Netherlands

thehyve

We empower scientists by building on open source software

How far does data travel?

During drug discovery, data generation begins with the identification of targets and finishes, often **over a decade later**, as a submission to the FDA[1]. Between discovery and submission, **data has to effectively move across disciplines, research activities and laboratories**. However, data often exists in disparate silos based on the owner and the question it aims to answer. For example, a pharma company can be organised based on disease areas, with each area having its own databases, registration systems and analysis tools. Across these areas, silos can extend as far as differing definitions of common concepts (fig. 1). Such disparate organizational methods may lend themselves to an individual or a group's workflow but hinder **universal querying**.

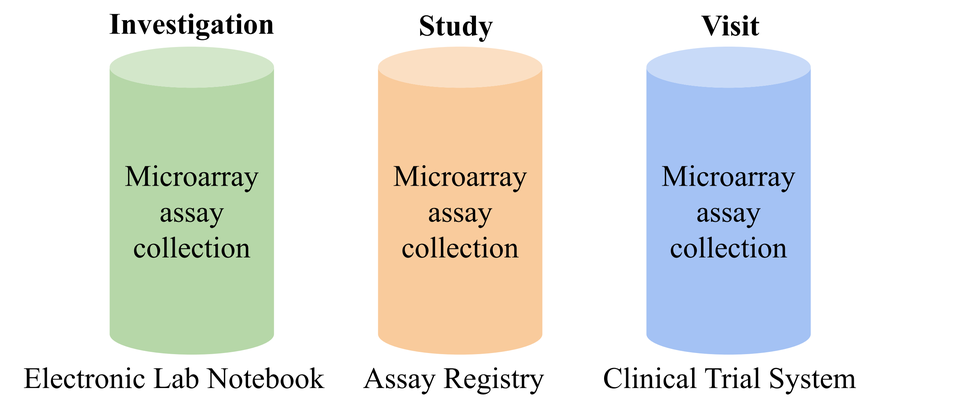

Fig. 1: A wet lab scientist records a set of microarray assays as an 'investigation' in a electronic lab notebook. Such a collection would be registered as a 'study'. Whereas, a set of assays is collected under the 'visit' term in a clinical study.

Linked data: connecting the silos

Connecting data allows for questions to be answered that individual silos cannot, without the need for costly technical and organisational change involved with abolishing silos. Ontologies serve as a starting point for developing a common understanding of concepts and the relationships between them. Where possible, the use of public ontologies establishes a common understanding within an organisation and beyond. Making data more Findable, Accessible, Interoperable and Reusable (FAIR)[2], brings value to every step in the drug discovery process, and linking concepts to public data can allow a researcher to perform an analysis against matched public data.

In Fig. 2, 8 sequencing assays are performed across 3 disease areas. The link between an assay and its disease area can be determined via the scientist who ran the assay. The assays are indirect instances of :genomicsAssay. If these triples were in a triple store, the following SPARQL query would retrieve counts of genomics assays across the organisation for each disease area:

```
PREFIX schema: <https://schema.org>
PREFIX dct: <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>
PREFIX rdfs: <https://www.w3.org/TR/rdf-schema/>

SELECT (count(distinct ?assay) as ?assay_count) ?disease_area where {
     ?assay a :genomicsAssay
     ?assay dct:creator ?scientist
     ?scientist schema:worksFor ?disease_area
}
GROUP BY ?disease_area
```

Open PHACTS[3] (Open Pharmacological Concept Triple Store), which was created to answer questions in drug discovery, is a great example of how integrating across ontologies makes it possible to derive implicit relationships between compounds, targets and pathways[4]. Another such example comes from researchers at Dalian University of Technology who built a knowledge graph by extracting triples from PubMed using literature mining. The method they developed was used to uncover associations between drugs and diseases by characterising the semantic links between drugs, targets and diseases[5].

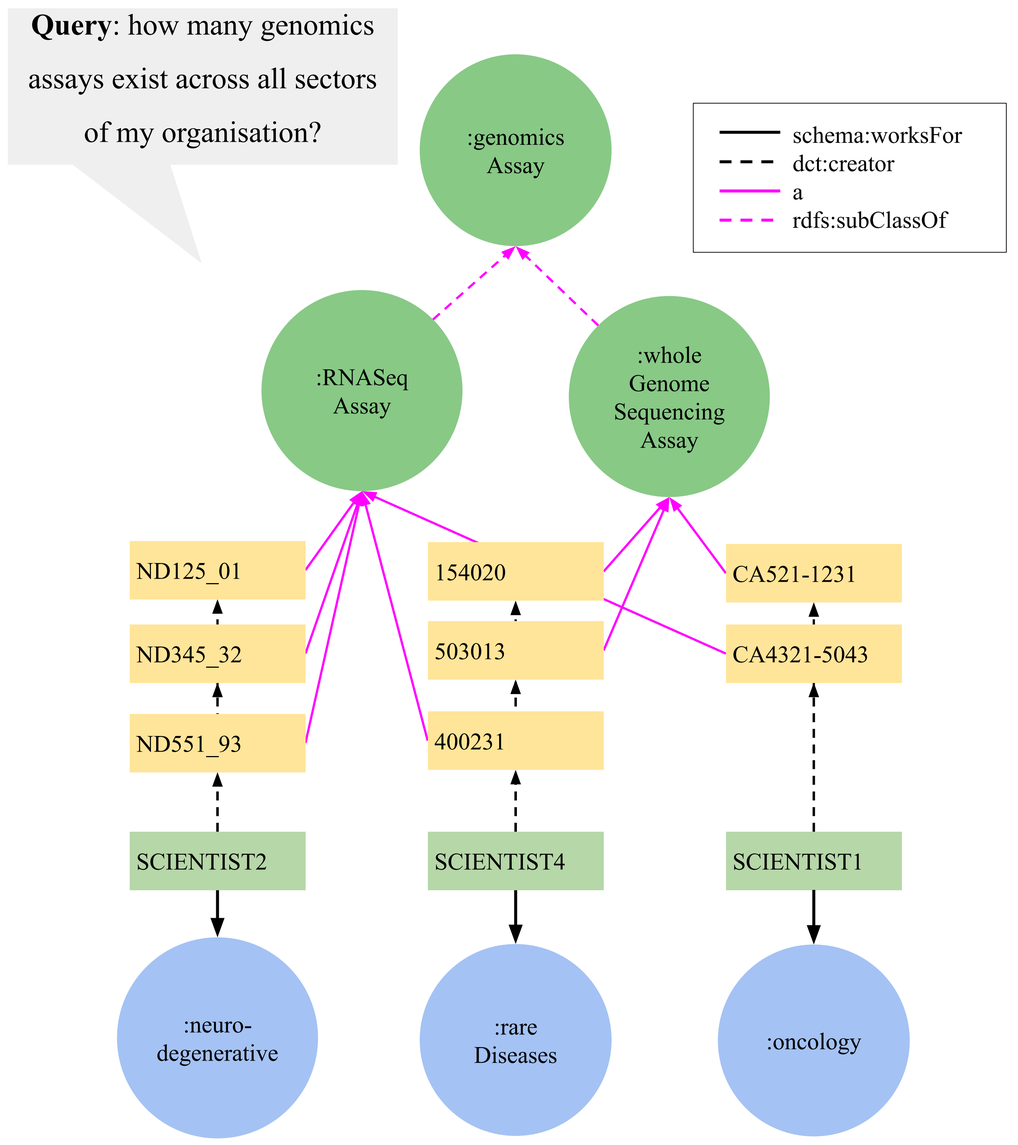

Fig. 2: Example ontology of 8 genomics assays performed by various scientists across 3 disease areas. External ontologies, schema.org, dcterms and rdfs, are leveraged. The property 'a' is shorthand for rdf:type.

Beyond ontologies: the power of inferencing

At The Hyve we have been developing semantic models for several of the top-10 pharma companies, to help them bridge silos and connect data from vastly disconnected areas of drug discovery. Recent challenges include linking concepts from external healthcare data to clinical trial data management and operations. We believe semantic data modelling along with proper (meta)data standards and a defined IRI schema, are fundamental to bridging the gaps across pre-existing data silos and adapt to a constantly changing data landscape. We continue to develop these standards and best practices for implementation of FAIR in pharma with our partners through projects like IMI FAIRplus and the Pistoia Alliance[6].

References

1. The Pharmaceutical Journal, PJ March 2015 online, online | URI: 20068196
2. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*3:160018 doi: 10.1038/sdata.2016.18 (2016).
3. https://www.openphactsfoundation.org
4. Samantha Kanza & Jeremy Graham Frey (2019) A new wave of innovation in Semantic web tools for drug discovery, Expert Opinion on Drug Discovery, 14:5,433-444, DOI: 10.1080/17460441.2019.1586880
5. Sang S, Yang Z, Wang L, et al. SemaTyP: a knowledge graph based literature mining method for drug discovery. BMC Bioinformatics. 2018;19:193.
6. Wise, John, et al. "Implementation and Relevance of FAIR Data Principles in Biopharmaceutical R&D." Drug Discovery Today, vol. 24, no. 4, 2019, pp. 933–938., doi:10.1016/j.drudis.2019.01.008.