Data Modeling and Governance in the Life Sciences

Rancho BioSciences

December 2021

In recent years, through a confluence of factors, we have witnessed a transformation that has led to the exciting era of 'Open Life Sciences Data' and resulted in an explosion in the availability of R&D data public resources.

Some of the factors include:

1) impact of the Internet and cloud infrastructure in "leveling the playing field" and increasing accessibility to providers, data, and tools
2) cost efficiencies bettering Moore's Law, has made Next Generation Sequencing and other science platforms available at lower price points (we already reached the $1,000 per genome price point[1, 2])
3) organizations have increasingly embraced open data principles over time.
    a. The Human Genome Project contributed with proactive steps with what has become known as the Bermuda Principles leading to immediate release and publication of sequencing assemblies throughout the project[3].
    b. The Supreme Court Molecular Pathology v. Myriad Genetics, Inc, case ruling that human genes could not be patented additionally led to increased open data[4].
    c. Additionally, as access and cost barriers have come down and as data standards have emerged, it has become easier to share data across the world. Many of the largest scientific data repositories now exist in the public domain[5]. Organizations such as the Broad Institute, the European Bioinformatics Institute (EMBL-EBI), the US National Center for Biotechnology Information (NCBI), the Swiss Institute of Bioinformatics (SIB), and many others have significantly contributed to open data principles through provision of excellent resources.
4) emerging analytical methods, including Artificial Intelligence and Machine & Deep Learning methods, require consistent high-quality data to deliver more reliable insights.
5) better understanding of biology and disease requires coordinated measurement and analysis of genomic, proteomic, cellular imaging, and many other of measures of disease and treatment.

Much of the data collected in the initial life science data wave were collected using FAIR-lite standards – standards that were mostly FAIR but left data scientists and informaticians wanting in various ways. While these resources provided a great leap forward for pan-omics platform data access and availability, data preparation time including collecting, cleaning, and organizing data still accounted for ~80% of a data scientist's total effort spent for data analysis efforts[6].

[1] https://www.forbes.com/sites/katiejennings/2020/10/28/how-human-genome-sequencing-went-from-1-billion-a-pop-to-under-1000/?sh=5ea851cd8cea
[2] https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
[3] https://en.wikipedia.org/wiki/Bermuda_Principles
[4] https://www.supremecourt.gov/opinions/12pdf/12-398_1b7d.pdf
[5] https://en.wikipedia.org/wiki/List_of_biological_databases
[6] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Opportunities: FAIR Data Principles

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Scientific Data*[7]. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data[8].

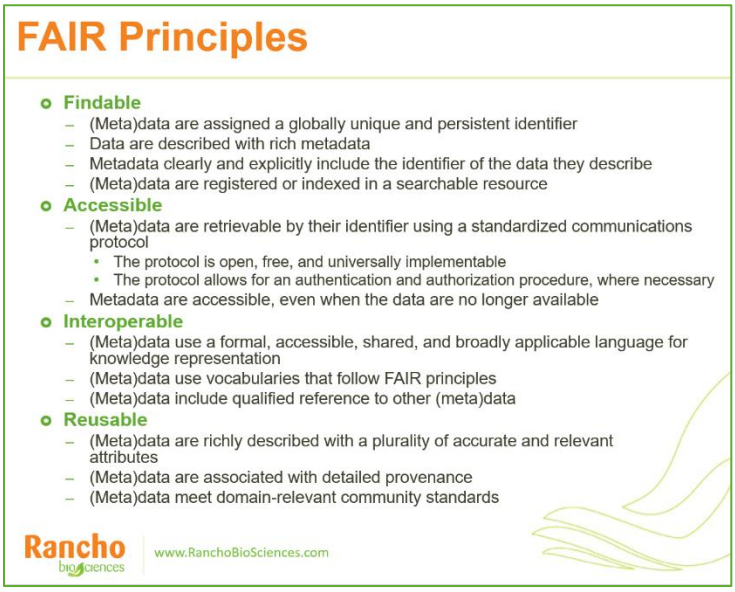

Figure 1: FAIR Data Guiding Principles

FAIR data principles were designed to help people align data, metadata, and infrastructure in a way that is logically searchable and usable:

Catching up with FAIRification and FAIR Data Standards

While FAIR data principles are designed to streamline access to R&D data, fully incorporating these principles across an organization is challenging. Multiple legacy data systems have often been built to local requirements, often for identical data types. Data systems are not designed to support accessible and interoperable data. Data and information provenance across internal systems is often not available. Interoperability with public data is complicated at best and often requires manual assessment and curation. Finally, the cost to retrofit existing solutions or implement new solutions that fully enable FAIR data standards is high and often not figured into IT solution budgets.

In short, most of the challenges faced by organizations looking to FAIRify their data and solutions, revolve around the absence or lack of completeness of broader, principled data modeling and governance practices. Implementing FAIR data practices across an organization requires significant change management for both IT and business practices. It requires substantive investment in data stewardship and new solutions. Most of all, implementing FAIR data practices requires robust data modeling and governance.

[7] https://www.nature.com/articles/sdata201618
[8] https://www.go-fair.org/fair-principles/

Driving Towards a Solution

Master Conceptual Model

The Master Conceptual Model is developed with the client to provide an authoritative overview of their overall business processes. It is typically a top-down modeling approach meant to support maximal breadth of organizational operations and goals. The master conceptual model is high-level, attempting to balance the twin requirements of being general enough to encompass all data domains while specific enough to be useful.

In the example shown in Figure 2, the Biologics Drug Discovery, Development, and Deployment Map (4DM) business process map from the NIH National Center for Advancing Translational Sciences is presented[9, 10].

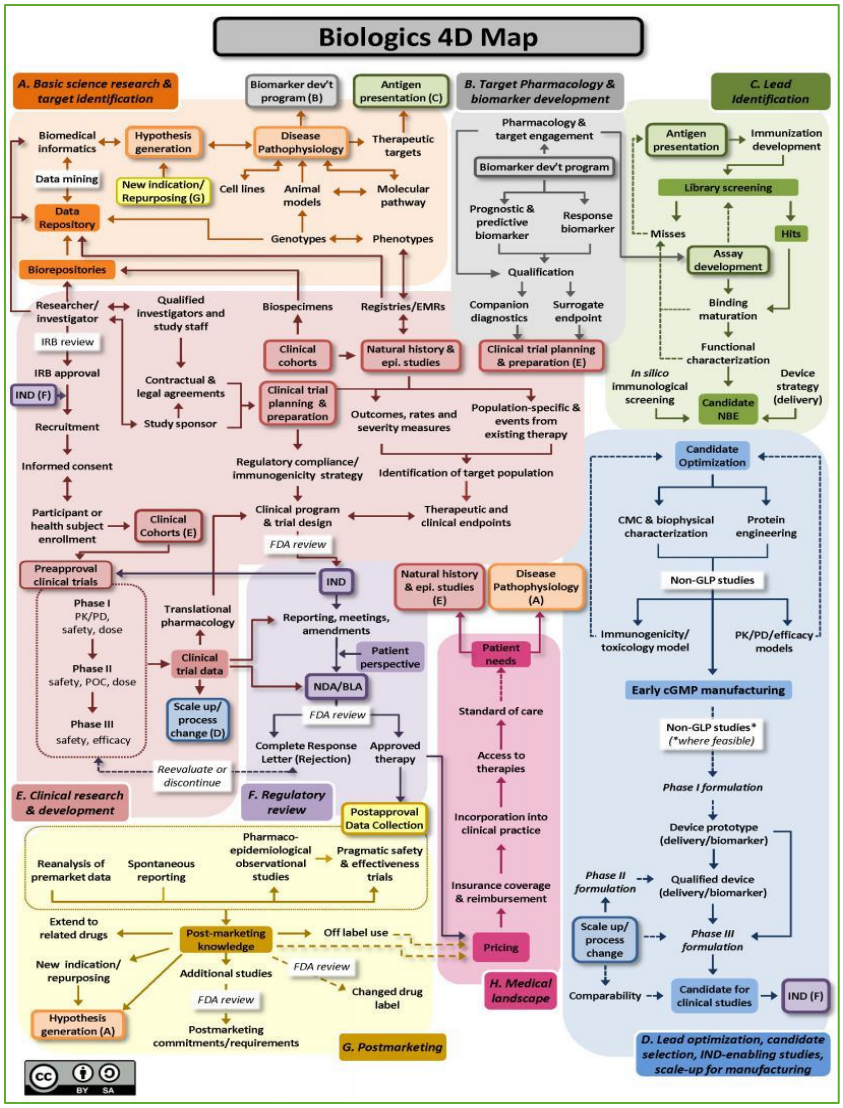

Figure 2: Biologics 4D (Drug Discovery, Development and Deployment) Map showing the basic entities and flows through the biologics theraeputic development process.

This business process map can, if developed through extensive business analysis efforts, be straightforwardly used to develop a master conceptual model. An example of key entities that support basic science research and target identification (Figure 2, top left) is presented in Figure 3. The master conceptual model shows key entities and relationships amongst them for capturing data and information related to animal model and cell lines studies. The same basic construct, because of the model's generality, can be used to describe human clinical trial studies required to support the clinical research and development business process shown at the middle left of Figure 2.

[9] https://ncats.nih.gov/translation/maps
[10] https://pubmed.ncbi.nlm.nih.gov/29269942/

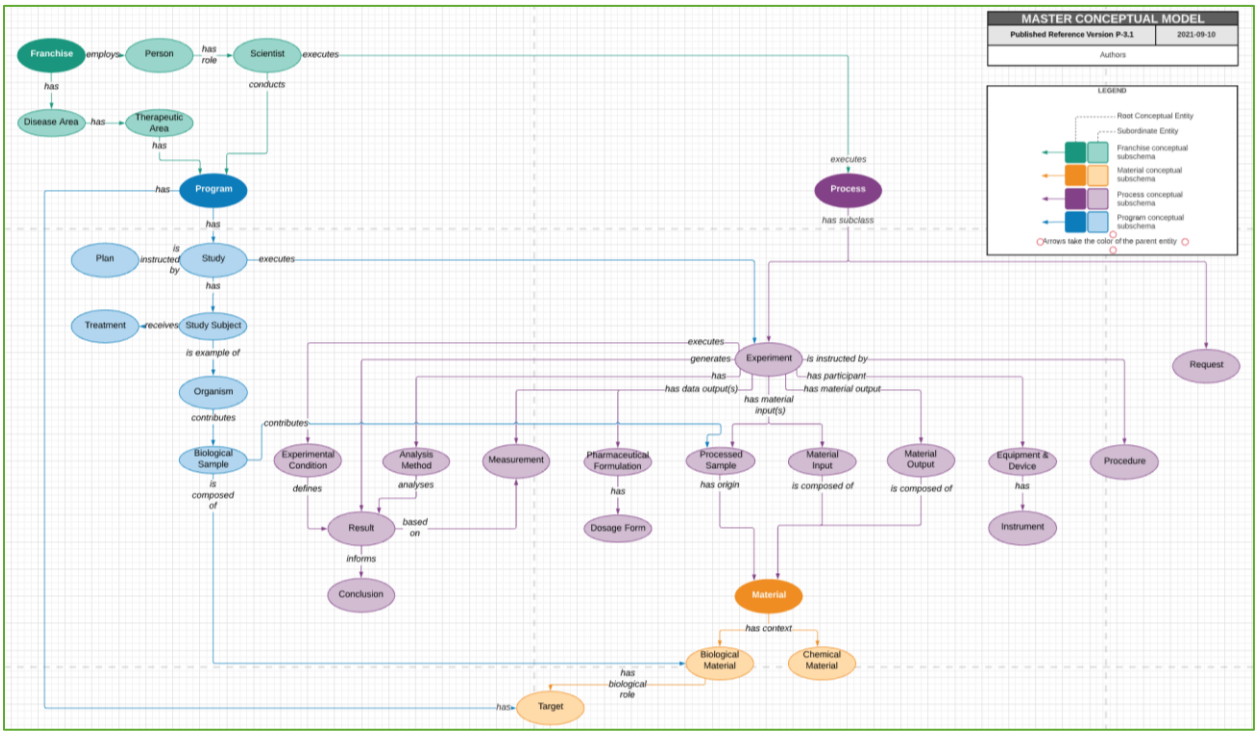

Figure 3: Master Conceptual Model for a basic Life Sciences Study and Experiment construct.

Contextual Models

Development and maintenance of Contextual Models follows the development of the master conceptual model. These contextual models provide clients with a framework for consistency through a detailed description of increasingly specific business practices. Contextual models inherit their base structure from the master conceptual model and allow for greater definition and differentiation of distinct domains, such as the research and clinical domains depicted in Figure 2. The entities in Figure 3 are more completely defined with entity subtypes and attribution. This continues through to the development of well-defined solution logical models.

The collection of the master conceptual model, business area contextual models, and solution logical models provides a roadmap of an organization's data and information. Combined with external data models such as FDAs Electronic Common Technical Document (eCTD)[11], the Observational Medical Outcomes Partnership (OMOP) common data model[12], the NIH Genomics Data Commons (GDC) Data Model[13], and others, an organization has not only their own data and information roadmaps, but also those for public resources. Alignment of internal and external models allows for simpler data exchange and direct data use operations. We are presently leveraging existing models, solutions, and application programming interfaces (APIs) and extending best practices where needed to enable our client's solutions and their access to public data resources.


```
{
    "tables": [
        {
            "table_name": "master_model_sample",
            "primary_key_field_name": "sample_identifier",
            "table_fields": [
                {
                    "field_name": "sample_identifier",
                    "ref_table": "master_model_subject",
                    "ref_table_key": "subject_identifier",
                    "ref_type": "Many-to-one",
                    "field_type": "NUMBER",
                    "field_length": 16,
                    "field_definition": "unique database identifier",
                    "field_source": "database sequence",
                    "unique": true,
                    "nullable": false
                },
                {
                    "field_name": "sample_number",
                    "ref_table": "",
                    "ref_table_key": "",
                    "ref_type": "",
                    "field_type": "VARCHAR2",
                    "field_length": 512,
                    "field_definition": "sample identification number - typically from a sample barcode",
                    "field_source": "text",
                    "unique": false,
                    "nullable": false
                },
                {
                    "field_name": "sample_mass",
                    "ref_table": "",
                    "ref_table_key": "",
                    "ref_type": "",
                    "field_type": "BINARY_DOUBLE",
                    "field_length": ,
                    "field_definition": "mass numeric value - paired with sample_mass_units",
                    "field_source": "text",
                    "unique": false,
                    "nullable": true
                },
                {
                    "field_name": "sample_mass_units",
                    "ref_table": "",
                    "ref_table_key": "",
                    "ref_type": "",
                    "field_type": "TEXT",
                    "field_length": 65535,
                    "field_definition": "mass numeric value - paired with sample_mass",
                    "field_source": "Units of measurement ontology (UO): http://purl.obolibrary.org/obo/UO_0000000",
                    "unique": false,
                    "nullable": true
                },
```


Figure 4: A programmatically accessible data structure based on a contextual data model.

Ontologies, Taxonomies, and Master Dictionaries and Controlled Vocabularies

Fully documented model references, and controlled vocabularies, ontologies, and taxonomies that describe the models must be leveraged if already existing, developed where they are not, and maintained. These ontologies, taxonomies and vocabularies are what fill the instantiated solutions with metadata and data. Utilizing best practices for selection, use, and maintenance of ontologies such as those presented by Malone, *et al.*, with *Ten Simple Rules for Selecting a Bio-ontology*[14] is critical for robust model definitions and data descriptions as described by FAIR.

In addition, utilizing public ontology resources enables use of important ontology class mapping resources. For example, SNOMED CT, ICD-10, and the NCI thesaurus (NCIt) all provide ontologies and taxonomies for important clinical entities and attributes. Each resource provides some support for cross-referencing or mapping of their terms to those within other ontologies. In this way, if one organization has chosen ICD-10 as a primary source and another organization has chosen NCIt, the cross-references provided by each, create a data translation layer that facilitates data management and interoperability requirements when data from disparate solutions are brought together.

[11] https://www.fda.gov/drugs/electronic-submissions-cder/electronic-common-technical-document-ectd
[12] https://www.ohdsi.org/data-standardization/the-common-data-model/
[13] https://gdc.cancer.gov/developers/gdc-data-model
[14] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004743

Data and Solution Governance

Simply put, consistent governance enables the implementation and maintenance of data and information practices and solutions such as those described above. By thoughtfully linking models and dictionaries to metadata, data, and business processes, an organization lays a solid foundation through FAIR data practices to meet business and regulatory requirements.

Rancho BioSciences provides its clients with FAIR data governance capabilities, providing data modeling services, ontology management, and solution development capabilities supported by documented procedures and policies. Our scientists, business analysts, ontologists, and solution developers work with you to understand your FAIR data requirements and partner with you to provide the best solutions implementing across your private solution space, with vendor and open-source solutions, and through the development and implementation of custom-fit FAIR solutions by facilitating data accessibility and availability, data usability, data integrity, and data security.