12. *INVESTIGATING FAIR PRINCIPLES IN NATIONAL INSTITUTES OF HEALTH (NIH) SCIENTIFIC DATA REPOSITORIES*

Marcello Peixoto Bax[150]

12.1 INTRODUCTION

Data collection and analysis are essential for all the sciences, but They are especially important when it comes to biomedical or health sciences. As the world advances towards the personalized medicine and more speed and agility in the production of drugs, the understanding of data collected in clinical trials and other studies is important to speed up the advance of scientific research. Unfortunately, due to the multiplication of data and formats, the collection, organization, and dissemination of data is becoming increasingly difficult to be efficiently executed. In addition, understanding phenomena and proving a hypothesis in this field require large amounts of data, and few researchers have the resources and means to collect such an amount of information. Clinical trials are expensive, require non-trivial resources and can take years to complete, depending on the study. Obviously, as such, this type of data collection and acquisition is not readily available for most researchers. That is why the health field, and other knowledge domains, are moving towards the widespread sharing of these data through public or private *data centers* available to researchers.

Unfortunately, as a result of inadequate data management, successful data sharing initiatives are very few. Data management is "the main channel for knowledge discover and innovation", promoting data sharing and reuse in scientific communities (Wilkinson *et al.*, 2016). It became important to define a set of common principles that define what a "good" data management should be. These principles, which highlight the findability, accessibility, interoperability, and reuse capacity of datasets, are known as FAIR Guiding Principles. Such principles are increasingly considered a reference for *data centers* and are being used to evaluate and highlight the success of certain initiatives. Numerous publications discuss the adherence to FAIR principles as a way to illustrate the commitment to facilitate data sharing in their respective communities. Examples include *Immune Epitope Database* (Vita *et al.*, 2018), *DisGeNET Platform* (Piñero *et al.,* 2017), *BioSharing Portal* (Mcquilton *et al.,* 2016) and *Omics Discovery Index* (Perez-Riverol *et al.,* 2017).

There are several common problems that prevent data to being considered FAIR. First, few *datasets* can conform to each other; several are closely related, but data are not set the same way; therefore, they do not easily conform and cannot be integrated for analysis. Data harmonization requires the use of common or at least commensurable categories and units of measurement. Second, in the case of research involving different teams, the research coordinator (main investigator) generally knows more details on the structural nature or data collected (their

150 Post-doctorate in Information Science, School of Information Science SIC – UFMG, bax@ufmg.br

properties and relations), than he can convey cohesively as supplementary information that would accompany the data itself (metadata). Finally, if the amount of data is sufficiently voluminous, automated methods may be the only viable way to generate comprehensive and in-depth analysis of it. However, if data meanings are not explicitly formalized in such a way to be "machine-readable", there will be no automated method that can support this analysis. This is where the concept of semantic "lifting" of data or even of semantic "ingestion" of data in repositories comes in.

12.2 SEMANTIC *LIFTING* OF DATA

The semantic lifting of data is a process by which data are converted from their original tabular representation to CSVs files and/or relational tables, for an ontological formal representation that represents the "knowledge" in the structure of a graph of knowledge (Pan *et al.,* 2017). In this operation, data are not only converted to another format, but "elevated" to the level of "knowledge" as they are represented by ontological models based on description logics that explain their formal semantic. The process transforms data, originally with no explicit meaning, into data potentially interoperable in the semantic web (linked data) and treatable by computer. Data lifting is, therefore, important because it helps to fight all the problems aforementioned that are targets for the treatment of FAIR principles. Data are collected and re-structured in accessible format, guided by metadata and machine-readable. Data in this format can be widely released by the web for subsequent extraction of information and knowledge, preserving its original meaning.

Several data centers are working to increase the *FAIRness* of their data repositories. In some cases, it is done with the development or integration of software platforms that incorporate a process of semantic enrichment of data models. Something that can be achieved by representing the model, or part of it, with a formalism guided by ontologies (semantic lifting) and mapping it to a reference ontology. As it is a relatively recent phenomena, there is not a consensual method of performing the process of *lifting* and data intake. Therefore, it is important to understand how different organizations are trying to improve the state of the art in terms of "semantic data lifting" as a way for us to learn from each of these efforts.

12.3 DATA CENTERS FUNDED BY *NATIONAL INSTITUTES OF HEALTH*

The National Institutes of Health (NIH) funds hundreds of data centers in different health areas. Some of them have unique data sets for a specific domain, while others host several data sets in various NIH domains and agencies. Although the Institute encourages data centers to use specific domain repositories whenever possible, these repositories are not available for all data sets. When researchers cannot find a data center that maintains a repository for its subject or for data they generate, a general repository may be a useful site to share data. General repositories accept data regardless of type, content or disciplinary focus. NIH does not recommend a specific general repository, but it maintains a non-exhaustive list, provided as a guide to finding repositories. The list contains the following most known general repositories: *Dataverse, Dríade, Figshare, Mendeley Data, Open Science Framework, Vivli and Zenodo*. A comprehensive list of data centers funded by NIH for sharing data was created by

the US National Library of Medicine, where *Trans-NIH Biomedical informatics coordinating committee* (BMIC)[151] keeps another list with currently 66 data centers (by October 2017) that maintain domain-specific and open repositories funded by NIH. Another 31 domain specific supported repositories include those that have limitations in sending and/or accessing data (sensible data).

Studying all the 97 repositories from these data centers would not be possible. Understanding the technical resources of a data repository is not trivial and generally requires accessing at least some available data. Thus, initially, **about 10 repositories were studied**, whose descriptions stood out for their greater level of detail. These repositories were inspected regarding the technical capacities that differentiate them from the others. The research revealed that, actually, some of these repositories are hosted in software platforms developed by third-parties that contain data from various studies and different institutions. It generates an interesting dynamic in which some data centers create repositories and host their data, while others simply host data for institutions that are interested. Finally, three data centers were selected for a more detailed inspection of their repositories: **ImmPort**, **Synapse** and **NDA** (*National Data Archive*) from the National Institute of Mental Health (NIMH). These three centers were selected due to the possibility of data access. Each one of the centers contains at least some repositories that allow the public access to summarized data, at the very least. Having access to data allowed a deeper understanding of how they is stored and how they can be searched. Another reason they were selected was due to their usability. These platforms had data search mechanisms relatively simple. It should be noted, however, that many other sophisticated data centers exist in many countries, including Brazil, and that this analysis did not intend to exclude their relevant contributions. However, the time and resources restrictions of this research demanded that some platforms easier to access were considered.

12.4 EVALUATION CRITERIA

The analysis of these data centers pointed out four evaluation criteria: 1) how can data be found and searched/consulted? 2) Are they single-domain or multiple cross-domain data? 3) Is the data representation scheme-free or fixed/relational? 4) Does the repository semantically lift the data when ingesting it into a database? These criteria were selected because each one of them works as an indicator of the level of data adherence to FAIR principles. The first criterion aforementioned looks to meet certain findability and accessibility characteristics because the filtration/query resources demand that data are "described with rich metadata", "recorded or indexed in a searchable form" and "retrievable by their identifier using a communication protocol standard" (Wilkinson *et al.*, 2016). The domain or knowledge area of a repository, according to the criterion aforementioned, works as an indication of potential for reuse of data. Although FAIR principles indicate that reusable data "meet community standards relevant to the domain" (Wilkinson *et al.*, 2016), the platforms that can meet the researchers' needs in correlation to domains have the potential to facilitate the research. Data must "use a formal, accessible, shared and widely applicable language for knowledge representation" so that they can be considered interoperable (Wilkinson *et al.*, 2016). Thus, the repositories that have a specific scheme (fixed/relational) use it as a way to facilitate this interoperability; however, fixed schemes can limit researchers in their decisions on which data they can or cannot

151 https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

store. Finally, machine readability of data was emphasized by proponents of FAIR principles. As such, what we are calling "data lifting" provides inherent mechanisms to meet each one of these principles.

The following sections detail the specific characteristics of each one of the repositories of three data centers researched, regarding the four evaluation criteria aforementioned. The degree of compliance with the criteria can be considered a proxy to understand how FAIR principles are more widely considered by data centers.

12.5 ANALYSIS OF THE STUDIED REPOSITORIES

According to Byrd *et al.* (2020), whenever possible, scientific research data must be shared through domain-specific repositories, which use data widely used in a field. Such specific repositories are ideal data warehouses. They provide long-term access to data through the provision of persistent IDs, such as digital objects identifiers (DOI). They reduce research costs by making large collections of correlated data available in a single location, which can reduce redundant work and encourage the generation of new hypothesis from secondary analysis. Lastly, they allow data to be cited, making data scientists earn credit for sharing datasets. We analyzed two domain-specific repositories below.

12.5.1 Domain-specific repositories

12.5.1.1 *ImmPort*

ImmPort[152] is funded by NIH, focusing on "Bioinformatics for the future of Immunology". It is a "curation and distribution portal" that aims at providing immunological data sharing (Bhattacharya *et al*., 2018); it is "one of the biggest open repositories with curation" of human immunological data (Sansone; Cruse; Thorley, 2018). In its efforts on data curation, ImmPort elaborates guidelines and standards based on suggestions from the immunology research Community, maximizing data accessibility and interoperability of this community. The repository is composed of four components: **private data, shared data, data analysis** and **resources**. Data collected is selected in the private data component, eventually released through the shared data component. The data analysis component uses *Galaxy* tool to allow data analysis in the repository space itself. Galaxy makes the analysis and meta-analysis of cytometry data easier, which is the focus of the portal. Finally, in **resources,** information on ImmPort is gathered, its publications and tutorials.

Unlike other repositories, ImmPort uses ontologies as a way to annotate its data with common and agreed terms, including one Cellular ontology, one Disease ontology, one ontology for Biomedical Investigations, one ontology for Proteins and one for Vaccines. These ontologies were used in the elaboration of *ImmPort Data Model*, which details the variables stored in each table and the relation among them. When uploading data to the ImmPort, the data model provides a set of common terms to be used so that the annotation is consistent with the other data already in the repository. It is done through data upload models and a validation tool. The studies available

152 https://www.immport.org/

in the repository can be consulted through a basic Keywords Search or by applying filters that include metadata such as if the study were or wasn't a clinical trial, the type of study, the species researched, the type of biological sample and the type of clinical trial.

These search resources do not consider data in itself, only certain metadata provided currently the study is submitted to the repository, which is done through pre-defined templates. Thus, there is no way to consult data by filtering certain criteria in various studies simultaneously. In addition, to visualize data in itself, individual archives must be downloaded by the researcher. However, detailed metadata on stored studies is available directly on the website. Metadata is standardized through templates designed based on the data models.

ImmPort domain is strictly of immunology. It is noted that the platform was adapted specifically to receive research on immunology. The data model itself also has elements very specific to immunology. Although this rigidity is important when data adjust to the model, it limits the use of ImmPort by other research with domains that could cross with immunology. ImmPort also has a specific scheme through the data model and, therefore, it is clearly not schema-free. This limits researchers when they need to store data that does not exactly fit the predefined schema/model.

ImmPort has some degree of data lifting, although it is not exactly clear how meaningful and comprehensive it is. The repository provides to the researchers models to be used when formatting and sending its data and asks the researchers to validate these data in relation to existing models. This indicates that ImmPort aims at standardizing its data so that studies are compatible among them. However, as studies are downloaded file by file, it is not clear whether the portal uses these models to store them in a way that they can be combined, generating information and knowledge. In general, ImmPort shows certain interesting resources, but it's not clear how deeply it applies the lifting process to store data.

Considering specific domain data centers funded by NIH, in addition to ImmPort, *National Data Archive* is an infrastructure to host data repositories in the mental health domain.

12.5.1.2 *NIMH National Data Archive (NDA)*

Initially developed to integrate a set of research data repositories as the *National Database for Autism Research* (NDAR[153]) and Other three in mental health, "it became a platform to share data on mental health and other researchers. The platform has strict restrictions on data use, and the download requires the user to complete a Certification of Use signed by NIH. Although it limits the use of the platform, summary of data is available and can be consulted. NDA has branched out to include other aspects of mental health domain. The repositories included, in addition to NDAR, are *Research Domain Criteria Database* (RDoCdb), *National Database for Clinical Trials related to Mental Health* (NDCT) and NIH *MRI Repository* (PedsMRI). NDA is structured to meet the needs of specific research data on mental health. In addition, the restriction of its access makes it accessible predominantly to participants of communities in the mental health area.

153 https://nda.nih.gov/

NDA content is organized around the concept of "Globally Unique Identifier" (GUID), which works as a way to identify data of unique individuals (Dan *et al.,* 2018). GUIDs are generated by a tool that requires the researcher to enter specific personal identification information, which is then used to generate a hash code that solely represents the person in the data set. The same personal identification information will ensure that the same GUID is generated, therefore, if the same subject participates in several studies, it will not be duplicated in the system. It allows that all data is internally related to only one person, enabling the NDA to provide sophisticated queries for data extraction.

NDA has six query tools: general queries, data from labs, data from papers, data dictionary, query per concept and query per GUID. Each tool provides its own exclusive resources, which improves the process of data collection and analysis. The general query allows the researcher to select predefined fields to build a query. The results from this query are shown (along with the summary statistics) and the resulting data can be downloaded. In addition, the user can select which exact fields he wants to download, as well as from which source. This is unique because it means that a single query can generate results in all repositories using NDA to store their data (although certification of data use is necessary to download data from each repository). *Data from Labs* and *Data from Papers* tools look into information on NDA collections and NDA studies, respectively. Here, collections and studies can be filtered by different criteria and downloaded using the same download mechanism used by the General Query tool. It is crucial because it allows the researcher to select several collections or studies and extract specific structures, in accordance to the definitions in the Data Dictionary, only for download. The *Data Dictionary* tool allows the researcher to select "data structures" and e "data elements" directly from the data dictionary. This dictionary shows several attributes of each data structure and includes detailed information on its elements. Finally, the *query by Concept* tool allows query through "ontological concepts", according to definition by ASD *Phenotype Ontology*, using the same filtering resources and download available in the platform as a whole. It is important to note, however, that data is not stored using any type of ontological representation, the ontology is used as a filtering tool. In fact, the NDA approach "does not allow easy creation of an ontology, whether it among all data in clinical evaluations in the NDAR or among data in the NDAR and other lexicon" (Dan *et al.,* 2018).

Just as the ImmPort, the NDA uses a very specific scheme, according to the definition in its data dictionary. When a user submits any data set, it has to be validated according to the data dictionary; otherwise, it is not accepted in the system. This validation tool is publicly available for researchers and will warn the researcher about his errors that can be fixed in his data. In addition, all data sets must have a GUID, which restricts them to be related to a single subject (it makes sense for clinical data and mental health, but makes extensibility low among domains).

If the researcher needs a structure not defined by the data dictionary, he can send new definitions to the NDA *Help Desk* for eventual implementation. This means that even if the platform has a very specific scheme, such scheme is in a certain way open to changes and additions. However, this makes changes in the scheme take longer to be implemented because all maintenance if performed manually by NDA employees. This is also applicable to data upload, which generally takes 4 months to be publicly available on the platform. Up to this point, data remain in a private status so that NDA employees can review and ensure its quality.

Due to the rigid scheme and NDA validation tools, the data intake process can be done quickly. The query tolls available suggest that data stored in the NDA are transformed from their original upload status to a format in

which all data are stored around GUID. It allows the extraction of knowledge among studies, collections, and repositories in a way many other platforms cannot.

The capacity of selectively manipulate data for download creates many opportunities for exclusive analysis of data. Our investigation was unable to exactly clarify how these data are stored in the "backstage", but it was clear that all data are associated to a single GUID in all the platforms in a way that it can be easily looked into. This makes NDA different from many other repositories; however, there is still room for improvement when it comes to upload automation and data curation.

12.5.2 Archive repositories for general use

Both data centers examined hereto are domain-specific, however, in certain circumstances, particularly at the beginning of the development of a scientific data domain, they may not have specific repositories. In such cases, investigators can still choose to put data in archiving platforms for general use, such as Figshare or Zenodo, along with metadata that precisely describe the archives included and its format. For data that cannot be publicly shared due to privacy issues, the Synapse platform provides a similar archiving platform for general use that supports controlled access sharing (Byrd *et al.* 2020).

12.5.2.1 *SYNAPSE Platform*

Synapse[154] is an open-source software platform for researchers who can use it as a site to store and annotate their data. Different from ImmPort, it does not have requirements for data formatting, serving researchers who just want to store their data somewhere. Even so, many data centers funded by NIH use it as a repository, and the platform itself is funded by several institutes linked to NIH.

The platform allows its users to create personal workspaces, upload different archives, connect themselves through provenience relations, annotate archives for better finding, provide narrative for data, create digital object identifiers (DOI) for any resource and work collaboratively. To register as a Synapse, the user simply has to provide an email, and the download of public data and the creation of content become easily accessible. Note that to store data on human beings (once there are use restrictions for that), the researcher must go through a certification process.

Synapse can be operated through several methods, including Python and R, in addition to the traditional web interface. However, certain functionalities (as downloading a group of files) are available only via Python or through the command line. Each resource in the platform has an exclusive SynapseID and, therefore, can be retrieved. The user must use the web interface to determine the resource SynapseID, but once found, automated tools can conduct data analysis.

154 https://www.synapse.org/

Each project in Synapse stores its information in relational tables whose schemes are defined by the project owner. This makes a Project queryable using similar SQL language, but in a standard research/filter interface also available. In general, these queries are used to research specific files or multiple files that share a specific characteristic; however, the user must know the scheme to generate a successful query. Synapse supports two different structures: table views and file views. File views allow browsing the uploaded files, view and download them. However, queries cannot be executed in data themselves. A data table can be researched and queried. Queries can be extended by different repositories, as each resource hosted in Synapse has an exclusive ID; however, as table schemes are identified by the Project owner, there is no guarantee that a single query can be successfully extended by several repositories that use different schemes.

As Synapse only operates as a domain independent platform, there is no specific domain connecting all repositories. Any certified user can upload and store data in the platform, a researcher does not need to operate in any specific domain to have the site benefits. Synapse is schema-free in the sense that the user is responsible for deciding which scheme to use to store data. However, each resource in a project must follow a pre-define standard scheme so that the project is uploaded and queryable. If the scheme is not enough for a researcher, he can create his project with his scheme, but it results in separation of an existing repository from which he could be benefited.

Relational tables are primitive when it comes to data lifting, for their rigid structures limit the information that can be extracted. As such, Synapse is limited when it comes to its data lifting resources. However, Synapse has an interesting capacity that allows users to "track the history of analysis and communicate and share a sequence of processing stages". The user himself must define the provenance (preferably When loads or edits an archive). Otherwise, the provenance track is lost. In general, the opening of Synapse and its low entrance barrier make the platform widely accessible, but this freedom is at the cost of making standardized approaches of data exchange among repositories impossible.

12.6 FINAL CONSIDERATIONS

Obviously, data organization and availability for research in health are highly related to the amount of information and knowledge that can be extracted from them. That is why research groups are starting to develop and apply tools to better structure these data so They can be analyzed and shared more promptly. Often, these groups appeal to FAIR guiding principles as a reference to develop their functional resources of data sharing. Making data available in this way, following such principles, lowers the barriers inherent in successfully carrying out research in health, allowing researchers to use and apply it in their research, given that many times data is not collected by them. It also opens doors for cross-domain collaborative studies, a trend that has been steadily increasing recently.

From the current 99 repositories funded by NIH, some have resources that illustrate the movement towards a wider sharing of scientific data by researchers, especially when it comes to data lifting, which is an important foundation, so data can be considered FAIR. All three repositories analyzed present certain attributes that show their movement towards a deeper use of data lifting in their infrastructure, although they are still very limited in the wider adoption of this criterium.

As summarized in Table 1, these repositories were evaluated through four criteria – query structure, scientific domain of action, data representation scheme and data lifting – as a way to understand the level of sophistication towards meeting FAIR data and machine-readable criteria. Although each one of the analyzed repositories has its strengths and weaknesses, it is important to understand what these organizations are doing to improve their resources aimed at sharing data for future research, as a way to understand the information and knowledge in the health research as a whole.

Table 1 – Analytical-comparative synthesis of repositories evaluated

Repositories/Criteria	ImmPort	NDA	Synapse
Query structure	Through metadata terms and filters	Through data from labs, data from papers, data dictionaries and concepts and GUID	Through table and archive
Scientific domain	Immunology	Mental health	any
Data scheme	Repository-specific	Repository-specific	*Project-specific*
Data lifting	Limited	limited	limited

Source: the author.

REFERENCES

BHATTACHARYA, Sanchita *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. ***Sci Data,*** v. 5, n. 180015, 2018, p. 1-9. DOI: https://doi.org/10.1038/sdata.2018.15

BYRD, James Brian *et al*. Responsible, practical genomic data sharing that accelerates research. **Nature Reviews Genetics**, v. 21, 2020, p. 615-629. DOI: https://doi.org/10.1038/s41576-020-0257-5

DAN, Hall *et al.* Sharing Heterogeneous Data: The National Database for Autism Research. **Neuroinformatics, v.** 10, n. 4, oct., 2012. p.331–339. DOI: 10.1007/s12021-012-9151-4

MCQUILTON, Peter *et al.* BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. **Database (Oxford)**, v. 2016, p. baw075, jan., 2016. DOI: 10.1093/database/baw075

PAN, Jeff Z. *et al.* Exploiting Linked Data and Knowledge Graphs in Large Organizations. Switzerland: Springer International Publishing, 2017. 266 p. DOI: https://doi.org/10.1007/978-3-319-45654-6

PEREZ-RIVEROL, Yasset *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. **Nature biotechnology,** v. 35, n. 5, p. 406-409, may., 2017. DOI: https://doi.org/10.1038/nbt.3790

PIÑERO, Janet *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, **Nucleic Acids Research**, v. 45, n. D1, p. D833-D839, jan. 2017. DOI: 10.1093/nar/gkw943.

SANSONE, Susanna-Assunta; CRUSE, Patricia; THORLEY, Mark. High-quality science requires high-quality open data infrastructure. **Scientific Data,** v. 5, n. 180027, feb., 2018 . DOI: https://doi.org/10.1038/sdata.2018.27

VITA, Randi *et al.* FAIR principles and the IEDB: short-term improvements and a long-term vision of OBO-foundry mediated machine-actionable interoperability. **Database (Oxford)**, v. 2018, p. bax105, 1 jan. 2018. DOI: 10.1093/database/bax105

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 160018, 2016. p. 1-9. DOI: https://doi.org/10.1038/sdata.2016.18

How to cite this chapter: BAX, Marcello Peixoto. Investigating FAIR principles in national institutes of health (NIH) scientific data repositories. *In*: SALES, Luana Faria; VEIGA, Viviane Santos de Oliveira; VIDOTTI, Silvana Aparecida Borsetti Gregório; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **FAIR Principles Applied To Research Data Management**: Brazilian Experiences. Brasília, DF: Editora Ibict, 2024. cap. 12, p. 145-154. DOI: 10.22477/9788570131959.cap12.