Evolution of FAIR data Supporting Digital R&D

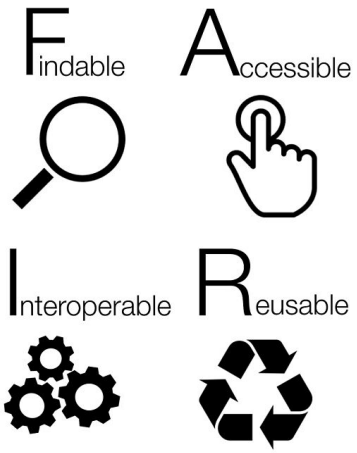
Curlew Research

Summary

- In the last few decades there has been an exponential expansion in the volume and variety of data generated by R&D in the Life Sciences.
- This expansion and the growth of “big data” has led to an increased urgency for organisations to better use these data for analysis & decision making both in the short term and into the future.
- The drive to better reusability of these data has led to the growth of the FAIR data principles; these aim to make scientific data Findable Accessible Interoperable and Reusable.
- Primary consumers of these large quantities of data are artificial intelligence (AI) and machine learning (ML) algorithms, but AI/ML based model building is only as good as the data inputs, thus reinforcing the need for FAIR.
- Making data FAIR will also help to address the growing reproducibility crisis across scientific research.
- Curlew Research has been at the forefront of data FAIRification, as this case study demonstrates.

Challenge

Modern R&D in life science research generates huge and ever-growing volumes of experimental and related data through the use of an ever-increasing number of technologies, especially those that operate in high throughput and high dimensions. Examples of such high volume, high dimension technologies include: biological and phenotypic screening; imaging; “omics”; and next generation sequencing related to genome & population studies. Historically, life sciences and healthcare data have been trapped in organisational and system silos. In many cases they continue to be trapped and this seriously hampers effective analysis, interpretation and reuse. With the rise of AI/ML within life sciences and healthcare, it has become obvious that a key blocker to success is not the maturity of the AI tools and techniques, but access to data of sufficient quality and in sufficient volume for the AI and ML methods to operate meaningfully. The phrase “no data, no AI/ML” is a frequently heard battle-cry of the current challenge. Even so, much of the data that is accessible has been created without due care and attention to reproducibility, quality, reusability and the FAIR[1] principles, which are now driving business improvement in data collection and annotation. This has led to another signature expression being heard frequently in data exploitation circles: “rubbish in - rubbish out”. Depending on the AI/ML model being developed, having access to a broad cohort of data from across the particular domain will be critical to ensure the necessary diversity, edge cases and breadth. It is this which will make the analyses successful and be broadly applicable: “quality in - quality out”. To remedy this, such data needs to be linked both internally and to external sources to make a FAIR data landscape which can power semantic models and knowledge graphs and so drive faster, better exploitation and decision-making.



Consequently, the scientific researcher of today faces an enormous challenge in fully exploiting all these valuable and possibly unstructured resources to derive new insights and drive scientific discovery.

[1] FAIR https://www.force11.org/group/fairgroup

Open PHACTS - An early FAIR project

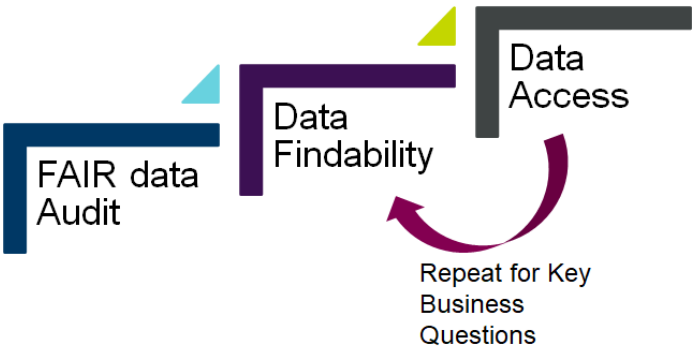
Open PHACTS[2] was a 5-year project of the Innovative Medicines Initiative (IMI), which started in 2010 with the aim of reducing the barriers to drug discovery in industry, academia and for small businesses. The resultant Open PHACTS consortium, which involved Curlew Research staff, built a freely available platform, integrating pharmacological data from a variety of information resources, and providing tools and services to query this integrated and interoperable data in support of pharmacological research. To further reduce the barriers to drug discovery, be it in industry, academia or for small businesses and start-ups, the Open PHACTS consortium developed the Open PHACTS Discovery Platform, which provides access to a collection of tools and services to query multiple integrated and publicly available data sources. The Open PHACTS Discovery Platform uses semantic technologies to provide a robust, adaptable framework for integration of multiple data sources into one coherent API. While the overall project had a predominantly pharmacological focus, it comprised a set of modular, reusable software components that could be used to address other scientific challenges.

One of the biggest challenges in the project was to enable interoperability between datasets and particularly using identifiers. We will describe the projects' approach to this challenge, with Curlew Research staff playing a leading role, as it provides a valuable case study reflecting on the importance of planning and having a FAIR strategy, and highlighting the criticality of Interoperability in order to make scientific data Reusable.

The Journey to FAIR data

The adoption of FAIR and the transformation of data is a journey driven by a number of factors including organisational pressures and needs, technical solutions and underpinning platforms.

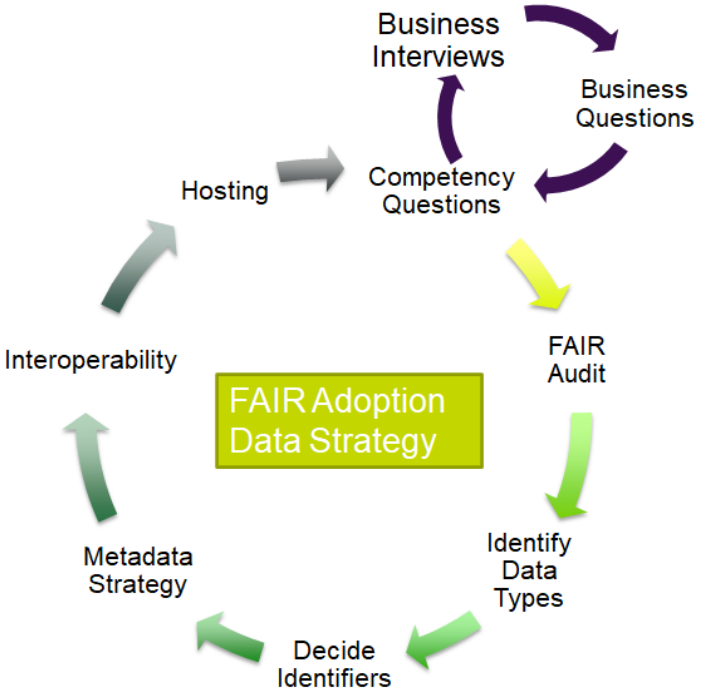
As operated in Open PHACTS, we believe that FAIR adoption should be broken down into several stages. The first phase, perhaps not surprisingly, should focus on the Findable and Accessible parts of FAIR. The reason for this is that starting the journey on the F&A will both support short term goals of where datasets are, and also begin the process for the harder parts of FAIR around Interoperability. Focusing initially on F&A will also deliver early value within discreet, already interoperable domains so providing support and justification for the ongoing investment in FAIR more broadly as part of the longer term strategy.



From our experience of FAIR projects, e.g. Open PHACTS, we would therefore recommend the following important initial steps:

- Identify the really key business questions that you are seeking to support within the organisation through extensive business interviews.
- Perform an initial FAIR data audit of your key systems and data.
- Review your capability maturity in the key dimensions of the audit.
- Begin your FAIR adoption journey in earnest - see below

[2] https://www.openphactsfoundation.org/



Below, we go into a little more detail on three of the absolutely critical steps in this adoption cycle:-

- Identify key business and core competency questions aligned to relevant business value.
    - This early analysis phase is critical to help focus and prioritise the initial work and phases longer term. In order to support meaningful progress, we believe it is essential to start with the business questions. This then guides the selection of the relevant data types and sources, which will help provide them to the data consumers in the required detail, format, quality and speed.
- Focus the FAIR audit on the data and business domain to support the business questions.
    - Focus on the data sources for the domain and consider the structure, format and demands. Do this in concert with the business experts, e.g. the data scientists, data consumers & creators. Such a FAIR audit is invaluable and can give early indicators for the challenge ahead based on the fundamental FAIRness of the data. A traffic light, red-amber-green system to indicate “as-is” FAIRness of current datasets and systems can show the extent of the journey that awaits and can help manage expectations amongst senior stakeholders. During this phase, key datasets can start to be listed in the chosen data catalog system. The data catalog helps to achieve the first part of FAIR - enabling the data to be Findable.
- Implement Identifier strategy
    - A key output of the FAIR audit will be the state of the identifiers used across the organisation and the use of external data sources and their identifiers. Experience of ontologies will be critical to support the implementation of relevant identifiers to enable Interoperability. The organisation will also need to be clear how it handles assets that only have internal identifiers and the ability to support mappings where appropriate.

Maintaining engagement with your core community throughout your FAIR journey is vital, and there are many ways to achieve this. One suggestion, employed successfully by Curlew Research, is to run datathons[3] during the course of the programme.

At its heart, FAIR and FAIR implementation is about change. Good change management with an emphasis on the importance of influencing behaviours at the point of data creation, instilling the concept of "data born FAIR" will be critical to ensure the long term success of the organisation's FAIR journey.

Finally, once your FAIR data strategy has been delivered as per the adoption cycle above, don't overlook the ongoing support that will be needed into the future. There are a number of key aspects of FAIR data governance and sustainability that will need to be included in the overall journey. Make sure you plan for and embed these longer term components in your FAIR data delivery projects.

Conclusion and Outputs

At its heart FAIR implementation is a journey that will deliver enormous value to the organisation and the scientists within, but only if there is a clear process and strategy for the adoption of FAIR principles. There is a huge buzz and growing usage of the FAIR term but to ensure the best outcome it's important to use experienced groups to support your steps along the FAIR strategy. Curlew Research's experience in Open PHACTS and other FAIR projects can help you with your journey to FAIR.

For further information please get in touch

curlewresearch.com

info@curlewresearch.com

@curlewresearch

https://www.linkedin.com/company/curlew-research

[3] A datathon is a data-focused hackathon — given a dataset and a limited amount of time, participants are challenged to use their creativity and data science skills to build, test, and explore solutions.