doi: 10.5281/zenodo.3970745

Chemical Data in Life Sciences R&D and the FAIR Principles

By Gerd Blanke[1], Thomas Doerner[2], Nick Lynch[3]

[1] StructurePendium Technologies GmbH, Reulsbergweg 5, D-45257 Essen, Germany

[2] Dr. Thomas Doerner Discovery Informatics Information, Obertüllingen 109, D-79539 Lörrach, Germany

[3] Curlew Research Limited, 9 Theydon Ave, Woburn Sands, Milton Keynes MK17 8PN, United Kingdom

In recent years, there has been an increasing drive towards scientific data not only being utilized for the specific context for which it has been created, but that data is findable, accessible, interoperable, and reusable beyond, by both humans and machines. This paradigm has been captured and formalized in the FAIR principles[1,2]. A number of initiatives have been formed to help promote and implement the FAIR principles and to make scientific data FAIRer, for example the GO FAIR Initiative[3] and - for biomedical research - the FAIRplus project[4] and the Pistoia Alliance FAIR Implementation project[5]. Up to now, many of the FAIR activities have either been centered on the needs of academic research, such as the output from academic research as published in scientific journals articles, or on domains which observe both a high value of data and a large gap to to fulfilling the FAIR principles, for example clinical research, where data is often captured in an uncontrolled manner as free-text.

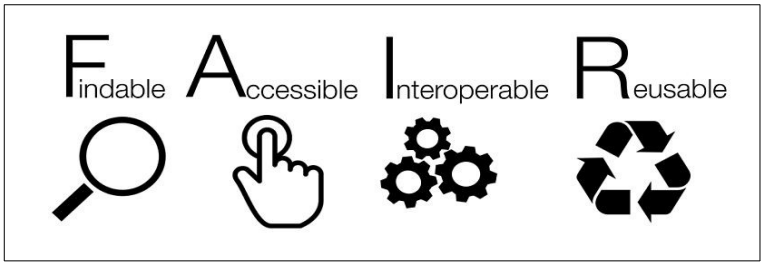

Figure 1: The FAIR principles (Source: https://en.wikipedia.org/wiki/File:FAIR_data_principles.jpg, author: SangyaPundir, license: CC BY-SA 4.0).

So far, chemical data such as chemical structures and chemical reactions and their associated data and metadata has not been a major topic of attention of the various FAIR initiatives. An attempt to introduce the FAIR principles to chemical data is the prototype of a FAIR Digital Object for the molecular structure, created by the Chemistry Implementation Network (ChIN)[6,7,8]. However, it mainly targets the needs of the academic research community, while the requirements in industry may be different.

Current State of Chemical Data in Industry

For company-internal data in life sciences and chemistry research and development, elaborated data management practices have been adhered to since many years. This in particular applies to chemical data, where chemical structures and chemical reactions are well-defined objects. Chemical registration systems and chemistry ELN's capture chemical compound and chemical reaction data as well as associated data in a

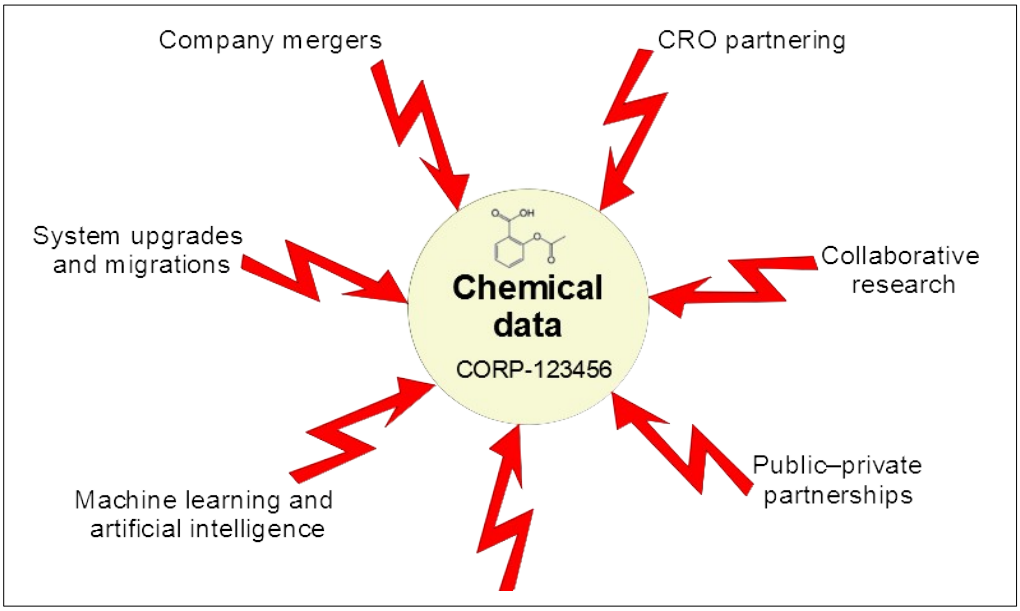

Figure 2: Stress on chemical data findability, accessibility, interoperability, and resuability.

structured manner, often supported by enforcement of detailed business rules and other data quality measures. Unique company ID's and chemical structure searching enable researchers to find, access and reuse existing data. Therefore, though many of the systems were created without the formal concept of FAIR data in mind, often company-internal chemical data already meets the findability and accessibility criteria, at least as long as you use of the data within the well-defined local environment the data has been created in. On the other hand, in particular chemical data interoperability has always been a challenge.

Challenges for Chemical Data FAIRness

The deficiencies in chemical data FAIRness become more apparent once the data is to be used outside its sheltered local 'silo'. Examples include combining chemical data from several systems, collaborations with CRO's, selecting external compounds to complement screening collections, public-private partnerships such as IMI[9] and Open PHACTS[10], and use of the data by autonomous devices and for machine learning / artificial intelligence. Another aspect revealing limitations is the ever increasing amount of change disrupting the current situation, which ranges from simple system migrations and upgrades via the general trend towards collaborative research up to large-scale company mergers and acquisitions.

Table 1: Examples of underlying issues affecting chemical data FAIRness. In particular, data interoperability poses a major challenge.

• Different chemical representations • Different chemical business rules • Missing, different, or non-standard taxonomies and ontologies for associated data and metadata • Identifiers that are not globally unique and that are only valid in the local context • Data locked in proprietary systems • 'Dirty' data

Lack of data FAIRness with respect to these wider contexts has significant impact on the business: it hampers data-driven innovation, jeopardizes the success of collaborations, and binds budget and resources in areas not contributing to the competitive edge.

The Path to Chemical Data FAIRness

In many cases, it will be a concrete business use case that drives the need for chemical data FAIRness. A recommended first step towards chemical data FAIRness is an assessment of the current situation. One should assess the local context, e.g. are there issues with findability or accessibility, are there interoperability issues, what hampers reuse of data, what is the situation for colleagues from other departments? You should also consider the wider company-internal context, e.g. what is the envisaged company-wide use of data, how do you ensure FAIRness along the whole data life-cycle? Subsequently, you should assess the external context, e.g. what data do you need to share externally, what are the deficiencies with respect to FAIRness of this data, and how do you ensure that incoming data is FAIR for you? What future collaboration situations do you envisage? Other aspects to consider include submission of data for publications, patents, or for registration at an authority such as FDA or REACH.

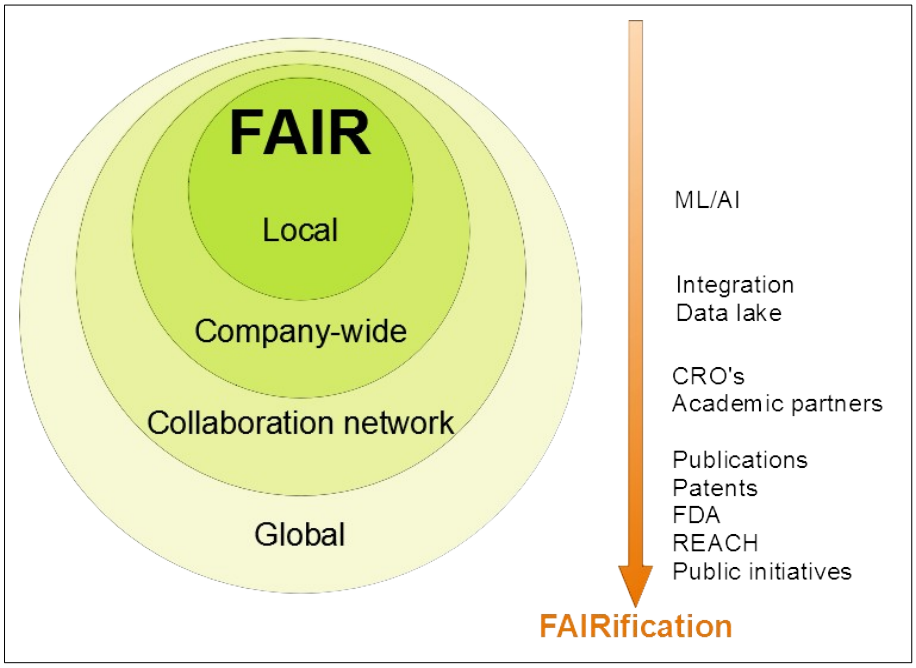

Figure 3: Different levels of FAIRness, driven by context.

While full FAIRification of existing data for use in a global context is definitely desirable, the effort will be substantial. Consequently, one might take a value-based approach that balances both actual use cases and associated requirements with the FAIRification efforts, and a pragmatic 'as-needed' approach with a proactive all-encompassing FAIR commandment. The same applies to your current data capturing systems. The situation will be somewhat different for new systems and for major system upgrades, where data FAIRness is easier to achieve and should be one of the fundamental considerations.

In parallel, the importance of data FAIRness needs to be looked at in a more holistic perspective independent of current use cases, under consideration of the whole data life cycle and with your strategic goals in mind. The output should be defined strategy for data FAIRness in your organization. This will allow to increasingly implement a FAIR-by-design paradigm for the whole data and system landscape.

Conclusion

In conclusion, a defined FAIR strategy for both existing and new data as well as for the associated systems can contribute to a competitive advantage in the quest for new biomedical modalities and to making your organization ready for the future. The authors can help assess current FAIRness of your chemical data, enhance findability, accessibility, interoperability, and reusability of it, prepare for future changes, and help remedy issues that you are currently facing. On a more strategic level, we can help shape an overall concept for data FAIRness and assist you in developing and implementing a FAIR-by-design strategy.

About the Authors

Gerd Blanke is founder of StructurePendium Technologies GmbH. StructurePendium offers consulting services in the area of chem- and bioinformatics with a major focus on standardization and normalization of chemical structures and reactions for registration and retrieval processes, e. g. in the context of database mergers, data transfers between different vendors, and data analytics.

Thomas Doerner is an independent specialist for Research Informatics in Life Sciences and chemistry. Located at the interface of R&D and Informatics, Thomas helps connect business needs, the user perspective and IT solutions to achieve outcomes.

Nick Lynch is founder of Curlew Research and interested in Life Science Data, AI & Informatics. Curlew Research works with Pharma/biotech and Life Science informatics/data science companies supporting their FAIR and data lifecycle needs.

Gerd Blanke, Thomas Doerner and Nick Lynch are members of the Informatics Alliance.

About the Informatics Alliance

The Informatics Alliance is a small group of dedicated chem- and bioinformatics experts focusing on serving the life science, agro and chemical industries. Each of us brings many years of experience with research informatics projects and practical implementations. We operate independently but we know and help each other, sharing experiences and expertise, and for bigger projects we join forces, for the benefit of all our clients.

References

Hyperlinks have been verified for correctness at the time of creating this document (July 2020).

1 https://en.wikipedia.org/wiki/FAIR_data
2 https://www.nature.com/articles/sdata201618
3 https://www.go-fair.org
4 https://fairplus-project.eu/
5 https://www.pistoiaalliance.org/projects/current-projects/fair-implementation/
6 https://www.go-fair.org/implementation-networks/overview/chemistryin/
7 https://www.mitpressjournals.org/doi/full/10.1162/dint_a_00035
8 https://github.com/easchultes/FAIR-Molecules
9 https://www.imi.europa.eu/
10 https://www.openphactsfoundation.org/