Related to other papers in this special issue	16 (p158); 14 (p139); 15 (p151)
Addressing FAIR principles	F, A, I, R

Taking FAIR on the ChIN: The Chemistry Implementation Network

Simon J. Coles[1†], Jeremy G. Frey[1], Egon L. Willighagen[2] & Stuart J. Chalk[3]

[1]School of Chemistry, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton SO17 1BJ, UK

[2]Department of Bioinformatics, Maastricht University, Maastricht, Limburg 6200 MD, The Netherlands

[3]Department of Chemistry, University of North Florida, Jacksonville, FL 32224-7699, USA

Keywords: Chemistry; Chemical information; Chemistry data; Chemical data standards; Infrastructure; Nomenclature; Molecular structure; Materials structure; Chemical reactions; Education; Community engagement; Endorsement and governance

Citation: S.J. Coles, J.G. Frey, E.L. Willighagen & S.J. Chalk. Taking FAIR on the ChIN. Data Intelligence 2(2020), 131–138. doi: 10.1162/dint_a_00035

ABSTRACT

The Chemistry Implementation Network (ChIN) is focused on supporting the FAIR Data needs of the research community regarding chemical related data. An Implementation Network is a consortium drawn from a community, in this case the chemistry discipline, committed to defining and constructing standards, materials and software in the spirit of the FAIR data principles and under the structure of the GO FAIR project. Furthermore, as a core science the ChIN has to reach beyond the chemistry community and support the use of chemical information in other disciplines. This will be facilitated through connections in the GO FAIR ecosystem of Implementation Networks. Examples of the FAIR chemical concepts that need to be supported include molecular and materials structures, chemical reactions, nomenclature and other chemical terminology and conventions. The ChIN aims to drive forward the application of the FAIR Data Principles relating to the full range of chemistry concepts that are key to the transparent and efficient communication of chemical information. Realizing the goal of FAIR chemistry data will require a culture change across the discipline. However this is best addressed once a critical mass of tools and approaches has been developed.

† Corresponding author: Simon Coles (E-mail: S.J.Coles@soton.ac.uk, ORCID: 0000-0001-8414-9272).

1. INTRODUCTION

Under the GO FAIR project, an Implementation Network [1] is a consortium drawn from a community that is committed to defining and constructing standards, materials and tools in the spirit of the FAIR data principles. The Chemistry Implementation Network (ChIN) [2] is focused on supporting the FAIR data needs of the research community generating and utilizing chemically-related data. More specifically information is regarding, but not limited to, molecular and materials structures, characterization data, chemical reactions, nomenclature and other chemical terminology and conventions. However, this work is broader than devising formats and standards and encompasses the full range of chemistry concepts that are key to the transparent and efficient communication of chemical information.

On the whole, the chemistry community does not have an inherent FAIR culture. While there are isolated exemplars, it is not the case for a large proportion of chemistry research and so the vast majority of data generated are not utilized. Moreover, chemistry is a subject that concentrates on exploring the landscape of possible chemical structures and relating them to physical properties. Additionally, there are also very powerful simulation techniques now being employed and so it is a discipline that could greatly benefit from machine learning methods to make new discoveries. However, to unleash this power it is necessary to move away from the current situation of databases, data silos and supporting information locked in journal PDFs and make more underlying data (both experimental and simulated) more available. Moving beyond these realms will not be possible without a FAIR approach.

2. FOCUS OF THE CHIN

The ChIN must therefore operate on a number of different levels, addressing a range of factors, which we have grouped as the following:

- *Organizational* – this aspect is about how the chemistry community (and interacting disciplines) effectively communicates the need for data and services for chemical information via the existing, accepted domain structure of international societies and professional organizations. This must also be appropriate for the different levels of resource requirements e.g., people, expertise, infrastructure;
- *Technological* – there is a requirement to understand and aggregate the software, code libraries, websites, and Web services necessary to support FAIR for chemistry-related data. It is therefore necessary to coordinate the "FAIRification" of existing services. These actions will subsequently enable the identification of gaps in service provision and therefore prompt targeted development in key areas. This focus is highly integral to the GO BUILD [3] pillar of GO FAIR;
- *Social* – addressing scientists' concerns about sharing data is a major hurdle in the move toward "data first science" and the broad community benefits it enables. ChIN will provide convincing use cases for how FAIR data accelerate science, and thereby will make scientists aware of available resources, services and training. This focus is highly integral to the GO CHANGE [4] pillar of GO FAIR; and
- *Educational* – the practices and principles developed need to be fundamentally embedded into the mindset of chemists and in particular the emergent generation. ChIN will steer educational guidelines and generate use cases and resources to enable the principles to be woven into the mindset of future generations of chemists. This focus is highly integral to the GO TRAIN [5] pillar of GO FAIR.

3. THE CRUCIAL ROLE OF COMMUNITY ORGANIZATIONS, UNIONS AND SOCIETIES

Chemistry is one of the core sciences – its community is vast in number and research diversity, reaching from environmental sciences, through biology to materials and physics to name but a few. So, there is not only a challenge to technically cover all these areas but also to reach out to the community and change working practice and social perspectives. In addressing the ChINs organizational focus, the role of learned societies and scientific unions is crucial in being able to achieve this, bringing both kudos and credibility and reaching to a wide range of researchers. The establishment of the ChIN fully recognized this need and Figure 1 illustrates the key partner organizations and their interoperation with the ChIN.

The International Union of Pure and Applied Chemistry (IUPAC) [6], through its Committee on Publications and Cheminformatics Data Standards [7] works with the ChIN, providing oversight and a mechanism to formalize outputs concerning standards and nomenclature that require international approval by the chemistry community. ChIN and IUPAC are complimented by work of the Chemistry Research Data Interest Group (CRDIG) [8], which operates as part of the Research Data Alliance (RDA) [9].

IUPAC, ChIN and RDA groups cannot work in isolation and therefore these three activities operate together in a coordinated fashion. The CRDIG group is concerned with technical development of standards, i.e., understanding the existing landscape of chemistry data standards and chemical data repositories, evaluating and updating existing standards and analyzing the need for domain specific repositories. ChIN acts both as advisory and technical development platform for the FAIRification of chemical data, both technically and on a policy level and focuses on the implementation of the outputs of the CRDIG. IUPAC is the international body for the chemistry community and acts with ChIN and RDA groups to formalize and endorse their outputs and then communicate them as widely as possible. IUPAC provides a governance structure for all of these activities.

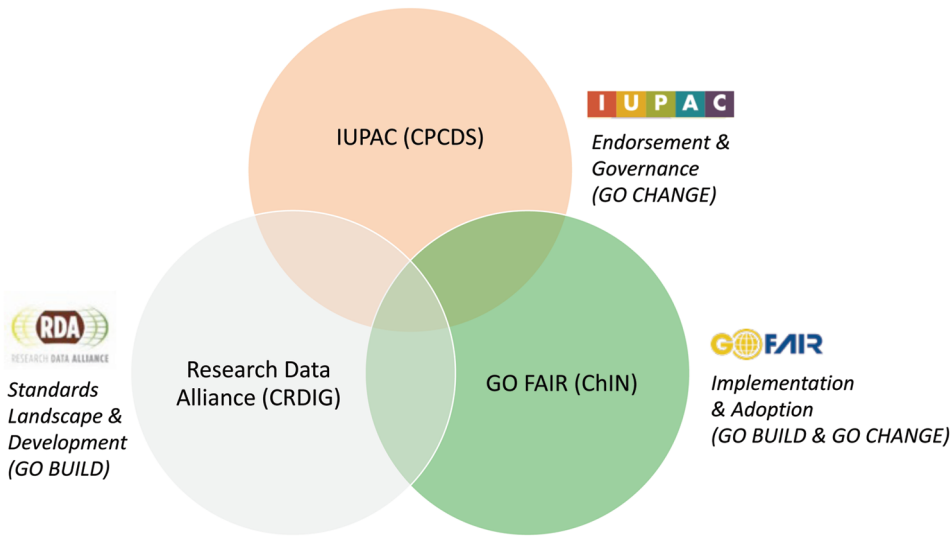

Figure 1. The interoperation between the ChIN and key community-leading organizations.

The development of the FAIR principles is fundamentally in alignment with the current focus of a "Digital IUPAC", which embraces the movement toward electronic (digital) access to chemical ideas, concepts and data for both human and machine use. For chemistry, this must go beyond conventional Open Access to journal articles and narratives and provide further access to chemistry-specific data, such as molecules and their properties, in a form that can be validated, used and reused.

Further engagement with the community has been achieved through workshops and publicizing to key organizations at key events. These include American Chemical Society (ACS) National Meetings, with the ACS Division of Chemical Information (CINF) [10], the Royal Society of Chemistry (RSC) Chemical Information and in Computer Applications Group (CICAG) [11], the IUPAC General Assembly, CODATA (International Data Week) [12], RDA Plenaries [13] and at the Beilstein Open Science [14] conference.

This awareness raising has led to a series of focussed technical workshops under the auspicious banner of IUPAC, co-branded with the ChIN. The workshops convene thought leaders, technical experts, active researchers, librarians, publishers and database providers with specific aims. These have included "Supporting FAIR exchange of chemical data through standards development" in July 2018 [15] and the NSF-funded "FAIR publishing guidelines for spectral data and chemical structures" in March 2019 [16].

4. BUILDING CHEMISTRY STANDARDS AND INFRASTRUCTURE

The domain of chemical sciences has a wide variety of data, which is at the core of the ChINs technological focus and forms part of the GO BUILD pillar of GO FAIR. Each of these data types requires specific best practices for which good exemplars, tools and resources need to be available. The following examples are key to the discipline and require work to better adhere to the FAIR Guiding Principles [17].

1). *Chemical identifiers*: Persistent, unique identifiers for chemical entities are required. For many circumstances these have been devised, but coordination and formalization is required to define the interplay required between identifiers for chemical structures, data, properties and reactions;
2). *Chemical data*: These are complex and may comprise a single data point, series of data points, or spectrum (array of data points). These need to be annotated with contextual information about the chemical system under study and the experimental conditions under which the data were collected;
3). *Chemical spectral files*: Data recorded on scientific instruments for characterization of chemical substances are a key type of chemical data and very widely used. They are not generally communicated in a findable format, but a well-used interchange format exists. Processes and protocols need to be assessed in light of the FAIR principles;
4). *Chemical structure visualisation*: There are numerous requirements for representation of a molecular structure, for which a variety of line notations, 2D and 3D representations exist and community agreement on standards is required;
5). *Chemical structure file formats*: Files or text identifiers that allow transfer of chemical structures between software applications critically underpin discoverability and communication of chemistry. A FAIR assessment of the many formats and their relationships is required;

6). *Chemical reaction visualization*: Representation of chemical reactions in 2D is the basis for communicating chemical transformations. De-facto approaches, driven mainly by the journal publication process, have been widely adopted and these need a FAIR assessment and formalization;
7). *Chemical reaction file formats*: Files or text identifiers that allow transfer of chemical reactions between software applications underpin chemistry communication in the digital age. Yet proprietary approaches tend to prevail here, but with numerous additional examples, such as the Reaction InChI (RInChI) [18], now becoming available but less widely used/accepted through open routes.
8). *Chemical terminology, properties, symbols and units*: The definition of meaning to chemical concepts has been a significant part of the effort of IUPAC in its 100-year history. Leveraging this knowledge for machine use will require community engagement in developing ontologies and standard ways to reference and represent physical/chemical properties and their associated symbols and units.

The above list comprises mainly standards and formats. However, for a fully functional digital environment for chemistry a more comprehensive infrastructure is required. The description of chemical concepts needs to be encoded digitally. Some aspects of chemistry have approaches to referencing concepts using vocabularies and ontologies. However, their use is limited and coverage of very sparse and more, coherent development at a community level is required.

The development, agreement and formalization of chemical conventions at a community level is required to bring together the above concepts and standards. This will be achieved by developing approaches (policies) in particular areas of activity e.g. standardizing chemistry in written communications, in recording in the laboratory and in interaction between machines.

The aggregation of these and other resources in the chemical space will enable us to work toward their FAIRification for the Internet of FAIR Data and Services (IFDS) implementation. On a technical level it will be necessary to integrate with, record workflows and generate FAIR digital research objects associated with these and here it will be instrumental to connect with the C2CAMP Implementation Network [19]. In order to record and map out the coverage of the FAIR chemistry data infrastructure, these chemical concepts and formats, along with other standards and policies, will be aggregated and analyzed in collaboration with FAIRSharing [20]. The FAIRsharing repository forms a key component of the FAIR StRePo Implementation Network [21] stores standards, policies and formats and has tools that enable gaps and overlaps with other activities to be identified.

5. CREATING A GO FAIR CHEMISTRY CULTURE

In order to address its social focus and in strong alignment with the GO FAIR pillar "GO CHANGE", the ChIN is establishing a working group focussing on implementation and community engagement and this will collaboratively (with CRDIG) promote FAIR through stakeholders and a group of recognized chemistry leaders who will champion by example. However, in order to affect change and promote development it is important to understand the current landscape and levels of FAIR principles implementation in the discipline. The map of existing FAIR chemistry resources, bodies and approaches generated with the

FAIRsharing organization will ensure that these are captured, curated and made visible to all. ChIN will then build on this map of resources to provide further ways to engage the community. For example, a current exploratory pilot project is underway to investigate the FAIR publishing of NMR spectroscopy data. This involves FAIR data generation in the laboratory (and the production of FAIR digital research objects as mentioned above) to propagate through to FAIR data publication with a traditional publishing organization.

As chemistry is a core science and chemical information is used widely in a variety of neighboring disciplines, the ChIN necessarily has to reach beyond the traditional chemistry community and support the use, and integration, of chemical information in other disciplines. Therefore, the work of the ChIN will benefit by drawing on the expertise of the GO Inter Implementation Network [22], which is concerned with linking and interoperability between disciplines. Technical interoperability will enable chemistry data that are generated or required by connected disciplines to be exchanged and used more widely. However, it will be necessary to ensure the same standards are adhered to and connections with other Implementation Networks in the GO FAIR ecosystem will greatly facilitate this process. These specifically include the currently active NOMAD Implementation Network [23] (concerned with materials discovery). However there are also close links with the preparatory GO NANO, Metabolomics and Geosciences Implementation Networks.

6. FUTURE DIRECTIONS

Ultimately the ChIN will act as a body to consult and to identify how organizations with chemical data can make their resources and services better aligned with the FAIR Data Principles and link suitably into the chemistry landscape of the Internet of FAIR Data and Services. This position will be built upon gaining recognition through a "ramping up" process, starting with key pilot projects based on existing resources and progressing through setting standards and facilitating larger, funded projects. It will facilitate development of approaches and policies and provide an oversight of the landscape of chemical information.

These efforts will not only drive convergence in chemistry but also with those disciplines around it. Therefore, the onus is on the ChIN to set guidelines that work at all levels i.e., organizational, technological, social and educational, and do not financially burden the community with the technological implementation of the guidelines.

To achieve this, the ChIN must build a body of FAIR resources and foster community engagement. The immediate priorities are to:

1). Save FAIR resources and services that are in peril i.e., orphaned through developers moving on, end of project funding, etc;
2). Aggregate existing FAIR resources and services (as identified through stakeholder engagement and work with FAIRsharing);
3). Facilitate the provision of FAIR services and resources e.g., by connecting people, projects and services.

4). Act as an advisory body for FAIR chemistry so that members of the community can approach ChIN and be signposted as to the appropriate course of action.

In the first instance ChIN will create a clear chemistry GO FAIR Web/digital/social media presence in order to make FAIR Chemistry more digestible to chemists. This will be done by creating chemistry-centric "personas" that provide guidance and links to FAIRsharing resources to all chemists (young and established), chemistry support staff and non-specialists wishing to use or create chemistry data.

The longer-term goals are around the ChINs educational focus and are to affect a culture change that moves the discipline community toward creators/authors producing FAIR data naturally as part of their routine workflows and ensuring that provision of FAIR data becomes a normal part of the formal publication process. This activity is closely aligned with the GO FAIR pillar "GO TRAIN". To achieve these goals, it will be necessary to promote the development of tools built on the foundations of FAIR principles and these must enable FAIR chemistry. It is crucial that such tools support researcher workflows both in the laboratory and the office, while providing added value and facilitating progress of the science. These tools would range from data capture in the laboratory e.g., Digital Research Notebooks, to authoring tools for communication and dissemination and ultimately to tools to exploit FAIR chemical data e.g., in Machine Learning approaches. It will also be key to encourage making more FAIR data available i.e., above and beyond formal publication, for example from PhD theses. The members of the ChIN and its collaborators will essentially be able to leverage existing projects and resources to fund future development at a larger, community-wide scale. ChIN will ultimately be considered successful if, empowered by the implementation of FAIR principles, and it can be said that the chemistry research discipline has been transformed by the move to integrating Data Science approaches into its everyday research practice.

AUTHOR CONTRIBUTIONS

S. Coles (s.j.coles@soton.ac.uk) coordinated, edited and contributed material, while S. Chalk (schalk@unf.edu), J. Frey (J.G.Frey@soton.ac.uk) and E. Willighagen (egon.willighagen@maastrichtuniversity.nl) all contributed material and reviewed content.

REFERENCES

[1] GO FAIR Implementation Networks. Available at: https://www.go-fair.org/implementation-networks/.
[2] The GO FAIR Chemistry Implementation Network. Available at: https://www.go-fair.org/implementation-networks/overview/chemistryin/.
[3] GO BUILD: FAIR technology. Available at: https://www.go-fair.org/go-fair-initiative/go-build/.
[4] GO CHANGE: Priorities, policies and incentives for implementing FAIR. Available at: https://www.go-fair.org/go-fair-initiative/go-train/.
[5] GO TRAIN: FAIR awareness and skills development training. Available at: https://www.go-fair.org/go-fair-initiative/go-train/.
[6] The International Union of Pure and Applied Chemistry. Available at: https://iupac.org/.

[7] IUPAC Committee on Publications and Cheminformatics Data Standards. Available at: https://iupac.org/who-we-are/committees/committee-details/?body_code=024.
[8] The Chemistry Research Data Interest Group. Available at: https://sites.google.com/view/digchem/.
[9] The Research Data Alliance. Available at: https://rd-alliance.org/.
[10] The American Chemical Society Division of Chemical Information. Available at: www.acscinf.org.
[11] The Royal Society of Chemistry Chemical Information and in Computer Applications Group. Available at: https://www.rsc.org/Membership/Networking/InterestGroups/CICAG/.
[12] International Data Week. Available at: http://www.internationaldataweek.org/.
[13] Future RDA plenary meetings. Available at: https://rd-alliance.org/plenaries.
[14] Beilstein Open Sciences Symposium 2019. Available at: https://www.beilstein-institut.de/en/symposia/open-science.
[15] Supporting FAIR exchange of chemical data through standards development. Available at: https://iupac.org/event/supporting-fair-exchange-chemical-data-standards-development/.
[16] FAIR publishing guidelines for spectral data and chemical structures. Available at: https://iupac.org/event/fair-publishing-guidelines-for-spectral-data-and-chemical-structures/.
[17] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.
[18] G. Grethe, J.M. Goodman & C.H.D. Allen. International chemical identifier for reactions (RInChI). Journal of Cheminformatics 5(2013), Article No. 45. doi: 10.1186/1758-2946-5-45.
[19] The C2CAMP Implementation Network. Available at: https://www.go-fair.org/implementation-networks/overview/c2camp/.
[20] The FAIRsharing project. Available at: https://fairsharing.org/.
[21] The GO FAIR making standards, repositories, and policies FAIR Implementation Network. Available at: https://www.go-fair.org/implementation-networks/overview/fair-strepo/.
[22] The GO Inter Implementation Network. Available at: https://www.go-fair.org/implementation-networks/overview/go-inter/.
[23] The NOMAD Implementation Network. Available at: https://www.go-fair.org/implementation-networks/overview/nomad/.