InChI As a Research Data Management Tool

by Henry S. Rzepa, Andrew Mclean, and Matthew J. Harvey

Progress in science has always been driven by data as a primary research output. This is especially true of the data-centric fields of molecular sciences. Scholarly journals in chemistry in the 19th century captured a (probably small) proportion of research data in printed journals, books, and compendia. The curation of this data from its origins in the 1880s and for most of the 20th century was largely driven by a few organisations as a commercial and proprietary activity. The online era, dating from around 1995, saw much experimentation centred around the presentation and delivery of journals, but less so of the data. The latter evolved, almost by accident, into what is now known as electronic supporting or supplemental information (SI), associated with journal articles. [1] That there was still a general problem in science was revealed by the "Climategate" events in 2009, where a lack of access to the data on which climate models are based induced all manner of unfortunate conspiracy theories. [2] These events catalysed a change in policy at, amongst others, UK research funders. One outcome of this change was seen in May 2015 with the introduction of new research data management (RDM) requirements for funded researchers. This centred around the precept that primary research data should be made openly available [3] and coincided with the evolution of the open science tripod of open data, open access articles, and open science notebooks. [4]

These new funder policies now require researchers to develop research data management plans, part of which involves publishing their data in what is called FAIR form. [5] The four components of FAIR are:

F: Findable. Data should be discoverable by searches, ideally on a global scale using consistent interfaces.
A: Accessible. Data should be openly retrievable not only by humans, but by machines operating on a larger scale for the purpose of data or content mining.
I: Interoperable. Once discovered and retrieved, data should be capable of validation and re-use, again not merely by human but also by software.
R: Reusable with a commensurate and declared license that allows this.

Although nowadays a virtually mandatory component of the journal publication process in chemistry, very little supporting information (SI) actually fulfils all these FAIR criteria for a variety of reasons. SI is mostly contained as a PDF document containing page breaks and page headers or footers. The PDF wrapper was never designed as a data container; such containment can easily disable data discoverability. Some data, such as crystallographic information, is contained in structured semantic form, but this is not generally true. Crucially, the PDF-based SI document never has formally declared metadata (information about the data contained therein) and its monolithic structure (examples have reached 504 pages in length, [6] and this may not have been even been close to the maximum) means that even a simple index of the text content is probably next to useless to satisfy the F of FAIR. SI is a child of its parent, the scientific journal article, and as such inherits the persistent (digital object) identifier or DOI of the article. The article DOI, however, carries no information (metadata) about the SI itself or about any data contained in the SI. The DOI normally points to a landing page for the article and this page has to be visually inspected by a human to ascertain the existence and whereabouts of SI, often in a manner parochial to the journal; a fail for both the F and the A of FAIR. Validation of data held inside a PDF file is rarely possible with any semantic assurance, a fail for the I of FAIR. Finally, the licenses that cover data are or should be fundamentally different from those that cover copyrightable materials such as journal articles. These are rarely declared; a fail for the R of FAIR.

All four aspects of FAIR can be addressed by the use of appropriately rich [7] metadata. In this regard, molecular science, and in particular molecule-centric chemical data, has been revolutionised by the introduction of the InChI identifier. [8] The key components and procedures for managing research data using InChI metadata identifiers include the following:

Table. Search queries enabled by the use of InChI identifiers in the management of research data in molecular science An HTML version of this table is available for download from http://dx.doi.org/10.14469/hpc/455

Search query [13]	**Description of the search**
http://search.datacite.org/ui?q=alternateIdentifier:InChIKey\:*	retrieves all instances where an InChI identifier is known
http://search.datacite.org/ui?q=alternateIdentifier:InChI\:*	retrieves all instances where an InChI key is known. The character "\" escapes the following character ":" to ensure it is part of the search string rather than the search syntax
http://search.datacite.org/ui?q=alternateIdentifier:InChIKey:CULPUXIDFLIQBT-UHFFFAOYSA-N	retrieves all instances where the InChI key CULPUXIDFLIQBT-UHFFFAOYSA-N is known.
http://search.datacite.org/ui?q=ORCID:0000-0002-8635-8390+alternateIdentifier:InChIKey\:*	retrieves all instances of a depositor with ORCID 0000-0002-8635-8390 AND (bolean) has an InChI key. The identity of the depositor can be resolved at http://orcid.org/0000-0002-8635-8390
http://search.datacite.org/ui?q=ORCID:0000-0002-8635-8390+alternateIdentifier:InChI\:InChI=1S\/C9H11N5O3*	retrieves all instances of a depositor with ORCID 0000-0002-8635-8390 AND (bolean) an InChI string that has the partial content 1S/C9H11N5O3 with the * representing one or more variable characters.
http://search.datacite.org/ui?q=has_media:true&fq=prefix:10.14469	retrieves all instances where any data media type is declared belonging to publisher 10.14469 (Imperial College)
http://search.datacite.org/ui?q=format:chemical/x-*	retrieves all instances where the data format type chemical/x-* [12] is declared
http://stats.datacite.org/?fq=datacentre_facet:"BL.IMPERIAL - Imperial College London"	retrieves resolution statistics belonging to publisher 10.14469 (Imperial College) for a selected month

- The SI document held on a publisher's site as part of a journal article can be augmented with or entirely replaced by the use of a data repository. [9]
- This repository should be capable of issuing an identified data depositor with a deposition receipt in the form of a DOI, issued by an associated authority. The current leading DOI registration agency for data is DataCite. [10]
- Such a DOI carries some assurance that metadata describing the deposition has been appropriately gathered and validated against a specified schema. In exchange for issuing a DOI, the issuing authority receives this metadata in a structured manner specified by a declared metadata schema and the entire process should ideally be automated as a workflow by the repository.
- The metadata schema includes core aspects such as the identity of the depositor (nowadays defined by their ORCID identifier), the data and time of the deposition, an explicit declaration of the license under which the data is issued, such as CC0, and the name of the publisher (normally the research institution).
- An InChI string and key for a molecule can be (automatically) generated and submitted to augment the core metadata, along with the media type of the data which greatly facilitates its semantic inter-operability.
- The registration authority in turn provides rich search facilities of the submitted metadata, along with access statistics.
- The registration authority can also record specific metadata specifying how the deposited data might be accessed based on its DOI, which allows implementation in a fully machine-automatable manner to allow high throughput access to data.

The data is now held in an optimal environment which includes appropriate metadata associated with a persistent identifier to ensure the data passes the FAIR

tests. Any journal article based on discourse or narrative where supporting evidence based on data is required can now simply include one or more data DOI citations in the bibliography. The article and data DOIs mutually complement each other. To show why InChI-based metadata in particular has the potential to catalyse enthusiastic adoption of RDM best-practices in molecular science, I will devote the rest of this article to a use-case example derived from our own experience and research.

A Use-Case Example

This research narrative, which has been peer reviewed and published in a journal, [11] describes the procedures and outcomes of curating a ten-year-old dataset of molecular files based on the NCI small molecule collection. The data and other research objects associated with this project were separately published in a data repository, cited in the bibliography of the article as refs 25, 27, 35, 36, and 50. It takes the form of an overall dataset collection assigned a DOI 10.14469/ch/2, with general metadata associated with the collection revealed using the query: http://data.datacite.org/10.14469/ch/2. There are 158,122 items within this collection (this is abnormally high, most collections would have far fewer items), each of which is also assigned its own DOI, *e.g.* http://doi.org/10.14469/ch/153690, and its own metadata; http://data.datacite.org/10.14469/ch/153690. Inspection of the metadata for any individual entry reveals the presence of both the InChI string and key as identifiers for the molecule in that entry, along with information about the media type(s) present for the data. For this dataset, the presence of a chemical/x-cml media type [12] suggests that a validatable XML-based document with implied identifiable semantic content present can be obtained: http://data.datacite.org/chemical/x-cml/10.14469/ch/153690. Such standardized metadata collection facilitates indexing, including that of the InChI identifiers, allowing a variety of rich searches based on it to be made (see Table). Both the search and the form of the outputs can be fully automated to allow high throughput queries.

Summary

As the management of research data together with its deposition as a digital research object becomes both increasingly common and likely mandatory, the deployment of rich metadata becomes essential. In molecule-based molecular sciences, the InChI identifier will play a pivotal role in enabling the discovery of the data and helping to ensure its FAIRness.

References

1. D. P. Martinsen, Supplemental Journal Article Materials in ACS Symposium Series, *Special Issues in Data Management*, 2012, Chapter 3, pp 31-45, doi: 10.1021/bk-2012-1110.ch003
2. O. Heffernan, *Nature*, 2010, 463, 860. doi: 10.1038/463860a
3. See https://www.epsrc.ac.uk/about/standards/researchdata/
4. C. L. Bird and J. G. Frey, *Chem. Soc. Rev.*, 2013, 42, 6754-6776. doi: 10.1039/c3cs60050e
5. See https://www.force11.org/group/fairgroup/fairprinciples
6. B. Bhaskararao and R. B. Sunoj, *J. Am. Chem. Soc.*, 2015, 137, 15712-15722. See the SI in doi: 10.1021/jacs.5b05902
7. The announcement of the detection of gravitational waves has associated FAIR data; doi: 10.7935/K5MW2F23 but the metadata (http://data.datacite.org/10.7935/K5MW2F23) cannot be described as rich.
8. The InChI identifier, see www.iupac.org/inchi
9. Research data repositories can be located using this resource: http://www.re3data.org
10. J. Neumann and J. Brase, J. Comp. *Aided Mol. Design*, 2014, 28, 1035-1041. doi: 10.1007/s10822-014-9776-5
11. M. J Harvey, N. J. Mason, A. McLean, P. Murray-Rust, H. S Rzepa, J. J. P. Stewart, *J. Cheminformatics*, 2015, 7:43. doi: 10.1186/s13321-015-0093-3
12. H. S. Rzepa, P. Murray-Rust and B. J. Whitaker, The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World-Wide Web information Exchange, *J. Chem. Inf. Comp. Sci.*, 1998, 38, 976-982. doi: 10.1021/ci9803233
13. A manual specifying the search syntax can be found at http://search.datacite.org/help.html

Henry S. Rzepa <h.rzepa@imperial.ac.uk> is Emeritus Professor of Computational Chemistry in the Department of Chemistry, Imperial College London.

Andrew Mclean is Research and Academic Support Team Leader in the ICT Division of Imperial College London

Matthew J. Harvey is a specialist in the High performance computing unit, ICT Division, Imperial College London