Metadata Recommendations for DataCite Registration: Draft recommendations for chemistry. V 1.0.2

Background

The increasing focus on the management of data and other research objects has identified the need to create appropriate infrastructures for handling metadata describing the attributes of data and file sets. The concept of FAIR data (F=findable, A=accessible, I=interoperable, R=re-usable) is closely associated with such metadata descriptors. DataCite is a so-called metadata registration agency operating on a global scale and the metadata is identified using a DOI (digital object identifier), a concept that has been almost universally used for this purpose in scholarly journal publishing and is now increasingly applied to research data publishing. DataCite also offer a valuable search interface which can be used to locate data based on its registered metadata. The core of metadata registration is defined by the schema used by DataCite, currently at version 4.1.[1] One schema element in particular is designed to be extensible by specific domains and this discussion document addresses how this element (Figure 1) might be exploited in chemistry.

Table 4: Expanded DataCite Recommended and Optional Properties

ID	*DataCite-Property*	*Occ*	*Definition*	*Allowed values, examples, other constraints*
6	Subject	0-n	Subject, keyword, classification code, or key phrase describing the resource.	Free text.
6.1	subjectScheme	0-1	The name of the subject scheme or classification code or authority if one is used.	Free text.
6.2	schemeURI	0-1	The URI of the subject identifier scheme.	Examples: http://id.loc.gov/authorities/subjects http://udcdata.info/
6.3	valueURI	0-1	The URI of the subject term.	Example(s) http://id.loc.gov/authorities/subjects/sh85026196 http://udcdata.info/037278

Figure 1. DataCite schema for the Subject property.

Applying the DataCite Schema.

The essence of applying a metadata schema is that it be integrated into a publishing workflow, with manual (human) entry deprecated to avoid errors and other inconsistencies. Such workflows are in turn best integrated into data publishing repositories. To facilitate such implementations, we have established[2] a pilot data repository which enables the schema to be tested and which also serves as an implementation test bed for recommendations.[3] Other well-known data repositories such as Zenodo and Figshare either do not enable rich exploitation of the DataCite scheme, require it to be populated manually or provide such features only *via* a commercially licensed version.

The subject schema (Figure 1) is implemented in XML as shown in scheme 1.

```
<subjects>
   <subject subjectScheme="Gibbs_Energy" schemeURI="https://doi.org/10.1351/goldbook.G02629"
   valueURI="http://gaussian.com/thermo/">-2861.844629</subject>
```

```
   <subject subjectScheme="inchi" schemeURI="http://www.inchi-
   trust.org/">InChI=1S/C24H32NOSi.C7H4ClN2O4.C6H15N.H2O/c1-5-6-19-25-20-13-18-23(25)24(26-
   27(2,3)4,21-14-9-7-10-15-21)22-16-11-8-12-17-22;8-4-5-1-2-6(9(11)12)3-7(5)10(13)14;1-5(2)7-
   6(3)4;/h5-12,14-17,19,23H,13,18,20H2,1-4H3;1-4H;5-7H,1-4H3;1H2/t23-;;;/m0.../s1</subject>
   <subject subjectScheme="inchikey" schemeURI="http://www.inchi-trust.org/">YAQIMGRXQRJPNV-
   AQUVTFJZSA-N</subject>
</subjects>
```

Scheme 1.

For each value of a subjectScheme, a SchemeURI is recommended and if possible it should also be a persistent identifier (PID) in the form of a DOI or other. In the above example, this DOI is provided by the appropriate entry in the IUPAC Goldbook if it exists. To disambiguate various possible implementations of a IUPAC-defined properties, a valueURI can also be provided. In this example, it is the implementation of the Gibbs energy thermodynamic property in the Gaussian program codebase. The URI shown here is not in the form of a persistent identifier. Where instances such as this are identified, it is suggested that the organisation owning the URL be contacted and urged to convert it to a PID such as a DOI.

One metadata is registered in such a form, it can now be queried using the DataCite interface. The (new) DataCite search engine is based on ElasticSearch, and the syntax (significantly different from the old syntax based on Solr which was superseded in January 2019) is shown in Scheme 2.

https://search.datacite.org/works?query=(subjects.subjectScheme:inchikey+AND+subjects.subject:KTOSDSJYNBIDCN-UHFFFAOYSA-N)+AND+(subjects.subjectScheme:Gibbs_Energy+AND+subjects.subject:"-1082.980914")

Scheme 2.

In plainer English, this translates to "find all registered instances of registered data for which the following two statements are both true; the subjectScheme is an InChKey with the value KTOSDSJYNBIDCN-UHFFFAOYSA-N **AND** the subjectScheme is a Gibbs_Energy with the value -1082.980914".

Recommendations resulting from this discussion paper.

In order to create a usable and consistent method of formulating search queries for data, the subjectScheme dictionary has to be clearly defined and potentially controlled so that it can be used when registering metadata with DataCite. Currently, the following subjectSchemes have been implemented and tested (as in Scheme 2) on the pilot repository.[4]

1. InChI
2. InChIKey
3. Gibbs_Energy

The priority for the DataCite working party is to formulate an extended list of SubjectScheme values, together with suggested options for the SchemeURI and valueURI (if needed), both specified if possible as PIDs. To facilitate implementation in the pilot data repository, the list should probably in the initial recommendation be limited to a small number (~10-20).

Organisation around specific types of dataset is probably desirable. In the first instance, it is suggested that NMR datasets be explored (Table 1).

Table 1. Suggestions for Subject extensions in the domain of NMR spectroscopy.

1. NMR_Nucleus (a controllable list)
2. NMR_Pulse sequence (should this be controlled?)
3. NMR_Solvent (should this be controlled?)
4. NMR_temperature (a numerical value, but declaring units requires schema additions)

A SchemeURI is suggested as 10.14469/hpc/4739 which is extensible depending on need. The IUPAC Goldbook only has an NMR entry (10.1351/goldbook.C01036, Chemical Shift) which may be too limited for our needs.

Workflows

For each recommended subjectScheme value, recommendations on how an automated workflow can be constructed should be suggested and ideally implemented in the testbed. For example, the existing workflow for InChI and InChIkey incorporated the opensource OpenBabel tool, which parses all uploaded datasets for identifiable molecular connection tables and derives the InChI string or key from them. For this purpose, at least one of the uploaded files has to contain this information. We currently recommend use of either a MDL molfile or a Chemdraw ..cdxml file for this purpose. For Gibbs_Energy it is a simple regular expression parser for the term in the Gaussian output log.

For NMR data, MestreLabs have kindly offered a tool based on existing NMR processing software, MestreNova which is already capable of capturing extensive metadata and which is being developed into a version suitable for inclusion in a suitable workflow. We hope to implement this tool in the workflows present in the Data Repository being used as a test bed by mid 2019 so that it can be tested by the community.

Implementation and Testing.

The existing testbed[2] can be extended to allow metadata key triples (subject, subjectScheme, value) to be selected by the depositor, which in turn will invoke an appropriate workflow. This can then be tested using queries modelled on that shown in scheme 1. A repository collection has been established with existing examples which can be added to.[4]

DataCite schema extensions.

As we develop and test our own metadata recommendations, we might wish to feedback suggestions to DataCite themselves for future features in their schema or tools. Three suggestions which have already emerged are;

1. The schema does not specify any method for data-typing. There is no way of specifying a floating-point value as being different to a pure alphanumeric string; both are treated equally at present.
2. The schema does not currently specify a means for declaring units for numerical values. For example, the one above is in units of Hartree, but there is no specification for declaring it.
3. The most serious limitation of the current DataCite indexing is that it does not honour grouping whereby sub-groups are separately evaluated. This can be illustrated using the variation on the search above; https://search.datacite.org/works?query=(subjects.subjectScheme:inchikey+AND+subjects.subject:"-

1082.980914")+AND+(subjects.subjectScheme:Gibbs_Energy+AND+subjects.subject:KTOSDSJYNBIDCN-UHFFFAOYSA-N)

If you evaluate it, you will notice it gives the same result as the original query, even though a value of -1082.980914 for an InChi string is clearly absurd. To solve this issue, we have persuaded DataCite to undertake a new re-indexing based on nested queries (https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-nested-query.html), scheduled to happen in February 2019. This will allow multiple <subject> entries to be correctly indexed and hence tested.

Requests for comment

The above is designed to promote discussion of recommendations for DataCite metadata registration. In particular we can decide if eg Table 1 is a useful initial template for creating a set of metadata descriptors which can be used to locate NMR datasets. Further potential subject areas can also be identified by the working group.

---

[1] https://schema.datacite.org/meta/kernel-4.1/

[2] M. J. Harvey, A. McLean, H. S. Rzepa, A metadata-driven approach to data repository design, *J. Cheminform*., **2017**, DOI:10.1186/s13321-017-0190-6, dataDOI:10.14469/hpc/1088

[3] A. Barba, S. Dominguez, C. Gomez, D. P. Martinsen, C. Romain,§ H. S. Rzepa and F. Seoane, A Workflow Enabling Facile Submission of NMR FIDs in Conjunction with NMR Spectra as Journal Article Supporting Information and FAIR-enabled data publication, *ACS Omega*, **2019**, in press. DOI:10.1021/acsomega.8b03005

[4] See DOI: 10.14469/hpc/4739