Harnessing FAIR data: A suggested useful persistent identifier (PID) for quantum chemical calculations.

Henry Rzepa

Published August 7, 2018

Citation

Rzepa, H. (2018). Harnessing FAIR data: A suggested useful persistent identifier (PID) for quantum chemical calculations. *Henry Rzepa's Blog*. https://doi.org/10.59350/nk414-18p76

Keywords

Interesting Chemistry, Academic Publishing, Chemical Context, Code, Data

Harnessing FAIR data is an event being held in London on September 3rd; no doubt all the speakers will espouse its virtues and speculate about how to realize its potential.♥ Admirable aspirations indeed. Capturing hearts and minds also needs lots of real life applications! Whilst assembling a forthcoming post on this blog, I realized I might have one nice application which also pushes the envelope a bit further, in a manner that I describe below.

The post I refer to above is about using quantum chemical calculations to chart possible mechanistic pathways for the reaction between a carboxylic acid and an amine to form an amide. The FAIR data for the entire project is collected at DOI: 10.14469/hpc/4598. Part of what makes it FAIR is the metadata not only collected about this data but also formally registered with the DataCite agency. Registration in turn enables Finding; it is this aspect I want to demonstrate here.

The metadata for the above DOI includes information such as;

1. The ORCID persistent identifier (PID) for the creator of the data (in this instance myself)
2. Date stamps for the original creation date and subsequent modifications.
3. A rights declaration, in this case the CC0 license which describes how the data can be re-used.
4. Related identifiers, in this case describing members of this collection.

The data itself is held in the members of the collection, each of which is described by a more specific set of metadata in addition to the more general types in the above list (e.g. 10.14469/hpc/4606).

1. One important additional metadata descriptor is the **ORE** locator (Object Re-use and Exchange, itself almost a synonym for FAIR). This allows a machine to deduce a direct path to the data file itself, and hence to retrieve it automatically if desired. It is important to note that the DOI itself (i.e. 10.14469/hpc/4606) points only to the "landing page" for the dataset, and does not necessarily describe the direct path to any specific file in the dataset. The ORE path can be used with *e.g.* software such as JSmol to directly load a molecule based only on its DOI. You can see an example of this here.
2. Each molecule-based dataset contains additional specific metadata relating to the molecule itself. For example this is how the InChiKey, an identifier specific to that molecule, is expressed in metadata;
   `<subject subjectScheme="inchikey" schemeURI="http://www.inchi-trust.org/">PVXKWVPAMVWJSQ-UHFFFAOYSA-N</subject>`
   The advantage of expressing the metadata in this way is that a general search of the type:
   `https://search.datacite.org/works?query=subjexts.subjectScheme:inchikey+AND+subjects.subject:CZABGBRSHXZJCF-UHFFFAOYSA-N`
   can be used to track down any molecule with metadata corresponding to the above InChIkey.

3. Here is more metadata, introduced in this blog. It relates to the (computed) value of the Gibbs energy (the energy unit is in Hartree†), as returned by the Gaussian program;
   `<subject subjectScheme="Gibbs_Energy" schemeURI="https://goldbook.iupac.org/html/G/G02629.html" valueURI="http://gaussian.com/thermo/">-649.732417</subject>`
   I here argue that it represents a unique identifier for a molecule calculation using the quantum mechanical procedures implemented in *e.g.* Gaussian. This identifier is different from the InChIkey, in that it can be truncated to provide different levels of information.
   - At the coarsest level, a search of the type
     https://search.datacite.org/works?query=subjects.subjectScheme:Gibbs_Energy+AND+subjects.subject:\-649.*
     should reveal all molecules with the same number of atoms and electrons whose Gibbs energy has been calculated, but not necessarily with the same InChI (*i.e.* they may be isomers, or transition states, etc). This level might be useful for revealing most (not necessarily all‡) molecules involved in say a reaction mechanism. It should also be insensitive to the program system used, since most quantum codes will return a value for the Gibbs energy if the same procedures have been used (*i.e.* DFT method, basis set, solvation model and dispersion correction) accurate to probably 0.01 Hartree.
   - The top level of precision however is high enough to almost certainly relate to a specific molecule and probably using a specific program;
     https://search.datacite.org/works?query=subjects.subjectScheme:Gibbs_Energy+AND+subjects.subject:\-649.732417
   - The searcher can experiment with different levels of precision to narrow or broaden the search.
   - I would also address the issue (before someone asks) of why I have used the Gibbs energy rather than the Total energy. Put simply, the Gibbs energy is far more useful in a chemical context. It can be used to relate the relative Gibbs energies of different isomers of the same molecule to *e.g.* the equilibrium constant that might be measured. Or the difference in Gibbs energies between a reactant and a transition state can be used to derive the free energy activation barrier for a reaction. The total energy is not so useful in such contexts, although of course it too could be added as a subject in the metadata above if a real use for it is found.
4. The searcher can also use Boolean combinations of metadata, such as specifying both the InChIKey and the Gibbs Energy, along with say the ORCID of the person who may have published the data;
   https://search.datacite.org/works?query=subjects.subjectScheme:Gibbs_Energy+AND+subjects.subject:\-649.*+AND+

> subjects.subjectScheme:inchikey+AND+subjects.subject:CZABGBRSHXZJCF-UHFFFAOYSA-N+AND+contributors.nameIdentifiers.nameIdentifier:*0000-0002-8635-8390♥

I have tried to show above how FAIR data implies some form of rich (registered) metadata. And how the metadata can be used to Find (the F in FAIR) data with very specific properties, thus Harnessing FAIR data.

---

†It is a current limitation of the V4.1 DataCite schema that there appears no way to specify the data type of the subject, including any units.

‡In theory, a range query of the type: `https://search.datacite.org/works?query=subjects.subjectScheme:Gibbs_energy+AND+subjects.subject:[\-649.1 TO \-649.8]` should be more specific, but I have not yet gotten it to work, probably because of the lack of data-typing means it is not recognised as a range of numeric values.

♥Implicit in this search is the grouping
https://search.datacite.org/works?query=(subjects.subjectScheme:Gibbs_Energy+AND+subjects.subject:\-649.*) +(subjects.subjectScheme:inchikey+AND+subjects.subject:CZABGBRSHXZJCF-UHFFFAOYSA-N)+AND+contributors.nameIdentifiers.nameIdentifier:*0000-0002-8635-8390
Currently however DataCite do not correctly honour this form of grouping.

♥Video of the speakers and the panel session at the end is now available.

https://orcid.org/0000-0002-8635-8390