IUCr forums

Discussions on IUCr projects and activities

Discussion of namespace conventions for community-developed CIF dictionaries

Snapshot of recent discussion thread, 14 August 2013

Formal proposal for privileged _audit.domain dataname

by **jamesrhester** » Tue Jan 29, 2013 7:31 am

To be submitted to COMCIFS together with the namespace policy. Please comment - I have speculatively added a policy for dataname scope at the end.

Proposal for a new dataname to support a CIF namespace mechanism

Background

We wish to build some sort of namespace mechanism into CIF so that other communities can use CIF with minimal, if any, coordination with COMCIFS. The key requirement is that datanames and the corresponding dictionary definitions must be unambiguously matchable. Currently, COMCIFS guarantees the uniqueness and immutable nature of datanames, so there is no need for any disambiguation mechanism. If CIF is to be usable outside COMCIFS, there must be a mechanism so that the readers and writers of CIF data files from a given community can agree on the correct definition for a given dataname.

Two partial solutions already exist:
(1) people and organisations register an opaque 'prefix' for a dataname with the IUCr. This allows users to populate their own namespaces safely and devolves management of dataname collisions to the relevant community. From the point of view of the outside discipline, there remains the annoyance that the datanames and dictionaries are cluttered with a redundant prefix.
(2) The _audit tags in a datablock can specify which dictionary the datanames come from. The problem then becomes one of encouraging programs to read and write these _audit items, given that simply finding a matching dataname in a datafile is already a pretty solid guarantee that it means what the programmer thought, as COMCIFS has up until now guaranteed the stability and uniqueness of datanames.

Some discussion has taken place in the namespaces forum and members are invited to read the comments there as well.

Proposed solution

We define an enumerated dataname, _audit.discipline, which takes values assigned by COMCIFS and should never be redefined by any CIF-using organisation - in effect it becomes part of the CIF specification. We can formally define a 'discipline' here as a collection of dictionaries which define datanames that are guaranteed to always have a constant, unambiguous meaning. This guarantee would presumably be provided by some organisation using policies chosen by that organisation. A CIF datafile wishing to explicitly specify which discipline its datanames are drawn from would set the value of _audit.discipline inside its datablocks. Likewise, programmers who are concerned about possible ambiguity in datanames can explicitly check for the value of this dataname.

Note the following:

- The IUCr would maintain a registry of accepted disciplines. In minimal form this could be the dictionary entries for _audit.discipline and something like _audit.discipline_URI
- There is no requirement to use the _audit.discipline dataname, nor to register disciplines. It is provided as a tool for

those wishing to avoid ambiguity

- Disciplines not wishing to register their discipline name but still wishing to use _audit.discipline, must never choose 'IUCr' (or whatever it is we decide) for their discipline name
- Minimal checking is required compared to the current _audit datanames, but similar guarantees of uniqueness and correctness are obtainable
- The _audit.discipline dataname should *never* be looped. Datanames drawn from multiple disciplines may not have overlapped when a datafile was produced, but may overlap when it is read, as there is no coordination between disciplines.

We further propose that the scope of the _audit.discipline dataname is the entire datablock and all save frames within that data block, unless a save frame gives a different value for _audit.discipline, in which case that new value will apply to all nested save frames within that save frame.