Integration of oreChem with the eCrystals repository for crystal structures

Mark Borkum, Simon Coles and Jeremy Frey
School of Chemistry
University of Southampton
{m.i.borkum | s.j.coles | j.g.frey}@soton.ac.uk

Abstract

This paper describes the integration of the oreChem Core Ontology (CO), a top-level ontology for the description of the planning and enactment of scientific methods, with the eCrystals repository for crystal structures. Records in the eCrystals repository constitute all fundamental and derived data that is obtained as the result of the execution of a crystal structure determination workflow. However, without a machine-readable description of the methodology, the provenance of data in the eCrystals repository cannot be determined. To facilitate the discovery and reuse of data in the correct context, we have described the eCrystals workflow using the CO.

1 Introduction

eCrystals - University of Southampton[1] is a repository for crystal structures generated by the EPSRC National Crystallography Service. The information provided by each record in the repository constitutes all fundamental and derived data that is obtained as the result of the enactment of a crystal structure determination workflow. The open availability of the data in eCrystals is intended to facilitate the verification and validation of the final crystal structures by independent parties.

The eCrystals workflow makes use of many specialist software applications, each of which can read and write data-files using a variety of crystallography formats. However, as the repository does not document the workflow that was used to generate each record, the provenance of the data-files contained by the record cannot be determined.

The benefits of exposing this data lie chiefly in reuse for computer assisted verification and validation of crystal structures. This work aims to make this data available in the correct context and in a semantically-rich form. In collaboration with members of the eCrystals team, we have developed and published a machine-readable representation of the eCrystals workflow using the oreChem Core Ontology (CO). We have also implemented software applications that leverage the oreChem metadata for each record to extend and enhance the functionality offered by the repository.

2 Preliminary

The oreChem Core Ontology (CO)[2] is a top-level ontology, specified in OWL DL [1]. The ontology provides terms for describing the planning and enactment of scientific methods, which are modelled as aggregations of linked information resources (referred to as "plan-things" and "run-things" respectively). Each run-thing is linked to exactly one plan-thing using a functional object property: `orechem:hasPlanThing`.

The methodology of a scientific experiment (referred to as a "plan") is modelled as an aggregation of operations and variables (referred to as "plan-stages" and "object-types" respectively), which are linked according to input/output (I/O) semantics. The methodology is 'realised' when it is enacted by a scientist. Each enactment (referred to as a "run") is modelled as a distinct aggregation of realised operations and variables (referred to as "stages" and "objects" respectively).

The CO uses SWRL [2] to infer the provenance of the objects that are realised during each run:

[1] eCrystals - University of Southampton – http://ecrystals.chem.soton.ac.uk
[2] oreChem Core Ontology – http://www.openarchives.org/2010/05/24-orechem-core-ns#

$$emitted\,(?\text{stage1}, ?\text{object}) \wedge used\,(?\text{stage2}, ?\text{object}) \Rightarrow followed\,(?\text{stage2}, ?\text{stage1}) \quad (1)$$

$$used\,(?\text{stage}, ?\text{object1}) \wedge emitted\,(?\text{stage}, ?\text{object2}) \Rightarrow derivedFrom\,(?\text{object2}, ?\text{object1}) \quad (2)$$

3 Method

The oreChem Core Ontology (CO) is used to describe the eCrystals workflow[3] as an instance of the `orechem:Plan` class. The planned execution of a software application is modelled as an instance of the `orechem:PlanStage` class. The inputs and outputs of each software application, i.e., the data-files, are modelled as instances of the `orechem:ObjectType` class and are linked using the `orechem:requires` and `orechem:emits` object properties.

An instance of the `orechem:Run` class is automatically generated for each record in the repository by comparing the list of available data-files with the list of object-types described by the plan. We use GraphViz[4] to visualise the instances of `orechem:Run` (shown in Figure 1). The rendered graphs are embedded into the eCrystals user interface as 'click-able' image-maps, which link back to the original data-files.

4 Conclusions

The eCrystals repository for crystal structures is an abundant source of experimental and derived data. This data is invaluable and of a high quality, however, without a machine-readable representation of the eCrystals methodology, is essentially unusable. The work described in this paper is dedicated to the semantic representation of the methodology of the eCrystals workflow, such that intermediate and derived data can be discovered and reused in the correct context.

5 Acknowledgements

The authors would like to acknowledge the support of Microsoft Research, who funded this research.

References

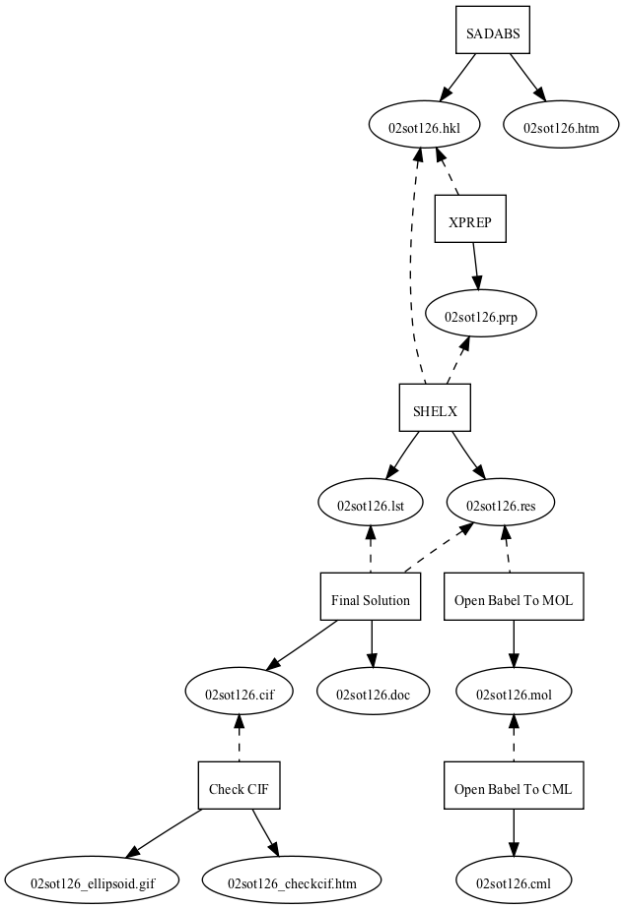
[1] D.L. McGuinness and F. Van Harmelen. OWL Web Ontology Language Overview. `http://www.w3.org/TR/2004/REC-owl-features-20040210/`, 10 February 2004.

[2] I. Horrocks, P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. `http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/`, 21 May 2004.

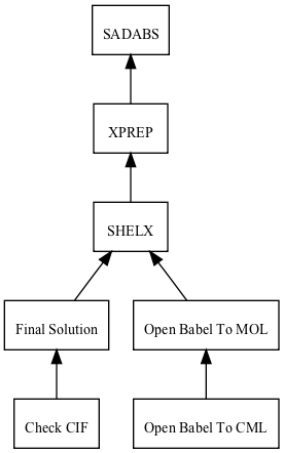
[3] G.L. Thomas, K. Navakhun, J.G. Frey, P.A. Gale, M.E. Light, S.J. Coles, and M.B. Hursthouse. bis(N,N'-bis(3,5-dinitrophenyl)isophthalamide)tetra-n-butylammoniumfluoride). `http://ecrystals.chem.soton.ac.uk/29/`, 23 October 2002.

[3] oreChem plan for eCrystals – http://ecrystals.chem.soton.ac.uk/plan.rdf

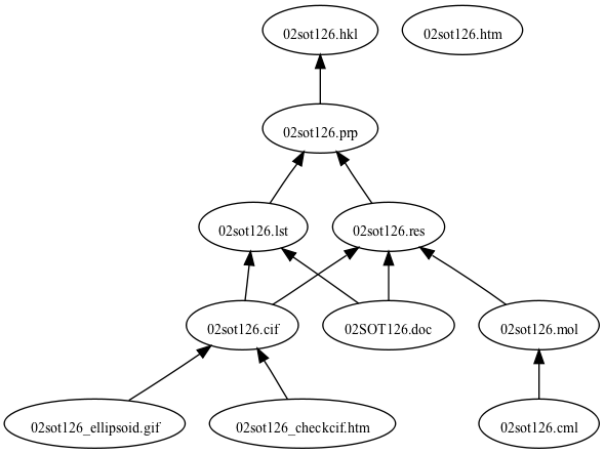
[4] GraphViz – http://www.graphviz.org



(a) Rectangles correspond to the execution of software applications (stages). Ellipses correspond to data files (objects). Solid and dashed edges correspond to assertions of the `orechem:emitted` and `orechem:used` predicates.



(b) Edges represent inferred assertions of the `orechem:followed` predicate.



(c) Edges represent inferred assertions of the `orechem:derivedFrom` predicate.

Figure 1. Inferred oreChem metadata for [3]