**PODD - An Ontology-centric Data Management System for Scientific Research**

**Yuan-Fang Li, Gavin Kennedy, Faith Davies,Jane Hunter**

The University of Queensland, Brisbane, Australia

{uqyli4, g.kennedy1, f.davies, j.hunter1}@uq.edu.au

Introduction

Data management has become a critical challenge faced by a wide array of scientific disciplines in which the provision of sound data management is pivotal to the success and impact of research projects. The huge and rapidly growing amounts of data to be managed and the fact that the models of data evolve over time contribute to making data management an increasingly complex undertaking that warrants a rethinking of its design. A number of intrinsic characteristics of Semantic Web ontology languages OWL [1] and RDF Schema [2], such as semantic rigor and the extensible nature, make them an ideal conceptual platform on which effective data management systems can be developed. We have designed PODD, a domain independent, ontology-centric architecture for data management systems that is open and extensible. In this architecture, the behaviors of domain concepts and objects are captured entirely by ontological entities, around which all data management tasks are carried out. Moreover, the open and semantic nature of ontology languages also makes such systems amenable to greater data reuse and interoperability. An ideal domain for applying these principles is phenomics, the systematic study of phenotypes of organisms. Phenomics research generates high volumes of heterogeneous data and makes use of emerging imaging and measurement technologies and processes, thus making it an ideal testbed for data management systems. In this context, we describe the development of the PODD[[1]](#footnote-1) data management system for phenomics research, as a step towards validating the practicality of the ontology-centric architecture.

Background

Data management is the practice of managing (digital) data and resources, encompassing a wide range of activities including acquisition, storage, retrieval, discovery, access control, publication, integration, curation and archival. For many data-intensive scientific disciplines such as life sciences and bioinformatics, sound data management informs and enables research and it has become an indispensable component [3]. The need for effective data management is, in a large part, due to the fact that huge amounts of digital data are being generated by modern instruments. Furthermore, the fast evolution of technologies/processes and discovery of new scientific knowledge require flexibility in handling dynamic data and models in data management systems.



## Figure 1: Attributes and relationships of domain concepts

Relational database systems have traditionally been used to manage research data, where database schemas are used as domain models to capture the attributes and relationships of domain concepts. In such a setting data mdel change is prohibitively expensive as schema redevelopment and migration are an error-prone and laborious task in practice. Hence, we believe that the traditional database approach is not suitable for domains where data and model evolution is the norm rather than the exception.

Semantic Web ontology languages such as RDF Schema and OWL possess expressive, rigorously-defined semantics and non-ambiguous syntaxes. Moreover, they have been designed to be open and extensible to support knowledge and data exchange on the Web scale [4]. These intrinsic characteristics make them an ideal conceptual platform on which a flexible data management system that supports dynamic data and models can be built.

The PODD Domain Ontology

Inspired by FuGe [5] and OBI[[2]](#footnote-2), we have created the base domain ontology in OWL to define essential domain concepts, their attributes and inter-relationships in an object-oriented fashion. As stated in the previous section, domain concepts will be modeled as OWL classes; relationships between concepts and object attributes will be modeled as OWL object and datatype properties. Concrete objects will be modeled as OWL individuals. Some high-level concepts and their attributes & relationships can be seen in Figure 1.

The ontology-centric architecture

The most distinguishing characteristic of the ontology-centric architecture is the central role ontologies play. In this architecture, raw data are not stored in a flat structure but are attached to domain objects organized in a logical, hierarchical way, defined according to the domain model that represents the structure of research activities. In this architecture, the behaviors of concepts and objects are all defined using OWL ontologies and RDF definitions and stored with them in the underlying repository. Complex relationships between different objects can be established, visualized and validated through these versioned ontological definitions. As a result, the architecture, as shown below in Figure 2, allows flexible addition, removal and modification of concepts and objects without affecting current objects.



**Figure 2: The high-level architecture of the PODD system.**

Based on the ontology-centric architecture, we have developed the PODD data management system to support Australian phenomics research. It supports data management tasks including data security, distributed storage, versioning, archival and publishing (making use of the *Identify My Data* service provided by ANDS). Moreover, it also provides RDF-based data exchange, reuse and discovery (e.g., query and search) capabilities and services.

Conclusion

The huge volumes of data and the fact that data models are constantly changing make sound data management practice a challenge faced by many data-intensive scientific disciplines. Traditional database-based systems are not flexible enough as the domain model is hardwired in database schemas and difficult to change. In this paper, we present an ontology-centric architecture where the domain model is captured in ontological terms, providing greater extensibility and is more amenable to model and data reuse and integration. As a validation of the feasibility of the architecture, we have also developed the PODD repository system based on this architecture as the core data management platform for the Australian phenomics research community.

References

1. Bechhofer, S., et al., *OWL Web Ontology Language Reference*, in *W3C Recommendations*, M. Dean and G. Schreiber, Editors. 2004, World Wide Web Consortium.

2. Brickley, D. and R.V. Guha, *RDF Vocabulary Description Language 1.0: RDF Schema*. 2004, W3C.

3. Gray, J., et al., *Scientific data management in the coming decade.* ACM SIGMOD Record, 2005. **34**(4): p. 34-41.

4. Auer, S., et al., *DBpedia: A Nucleus for a Web of Open Data*, in *Proceedings of the 6th International Semantic Web Conference (ISWC)*. 2008, Springer. p. 722-735.

5. Jones, A.R., et al., *The Functional Genomics Experiment model (FuGE): an Extensible Framework for Standards in Functional Genomics.* Nature Biotechnology, 2007. **25**(10): p. 1127-1133.

1. <http://www.itee.uq.edu.au/~eresearch/projects/podd/> [↑](#footnote-ref-1)
2. <http://purl.obolibrary.org/obo/obi> [↑](#footnote-ref-2)