# Using eXframe to build Semantic Web Genomics Repositories

Emily Merrill[1], Shannan Ho Sui[2], Stéphane Corlosquet [1], Tim Clark[1, 3, 4] and Sudeshna Das[1, 3]

[1] Massachusetts General Hospital, Partners Research Building, 65 Landsdowne St, Cambridge, MA, 02139, USA
[2]Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA
[3]Harvard Medical School, 25 Shattuck St, Boston, MA, 02115, USA
[4]School of Computer Science, University of Manchester, Oxford Road, Manchester, UK, M13 9PL

**Abstract**. We would to like to present eXframe: a software platform for developing Semantic Web genomics repositories. eXframe is implemented using Drupal 7, an open-source PHP/MySQL based content management system. eXframe provides a user-friendly interface for researchers to enter the information about their experiments and share these with their colleagues and external collaborators. An underlying Drupal-based content model, specialized for genomics, represents the provenance, assays, samples and data produced in the experiment. The relevant metadata fields in this model are mapped to established biomedical ontologies, which enable extended search across useful parameters such as experiment type, technology platform, model organism, authors, genes, proteins, and so forth. Using these mappings and the Drupal RDF modules, eXframe genomics data can be automatically published as Resource Description Framework (RDF) to produce Linked Data. The RDF is indexed to produce a SPARQL endpoint using the PHP ARC2 libraries. We will demonstrate how to use eXframe, create ontology mappings and run sample queries using the SPARQL endpoint.

**Keywords:** Drupal, Genomics, RDF, SPARQL

## 1 Introduction

There are only a handful of examples of genomics repositories available as Linked Data, but none of them are available as reusable systems. Recently, the Functional Genomics Production Team at the European Bioinformatics Institute (EBI) made their gene expression data from Expression Atlas [1] available as Linked Data that can be queried using a SPARQL Protocol and RDF Query Language (SPARQL) endpoint. These Semantic Web technologies allow flexible graph based querying as well integration with other ontologies or knowledge repositories.

We have developed a reusable platform for building genomics repositories, eXframe [2,3], which automatically formats the stored experimental data as RDF and indexes it into a SPARQL endpoint. The platform handles a variety of genomics data types including microarrays and next generation sequencing technologies such as RNA-Seq, ChIP-Seq, Bisulphite-Seq and RIP-Seq among others. In this presentation, we will present details of the model we developed, ontology mappings and sample SPARQL queries on a genomics repository.

## 2 Methods

We used the open source content management system, Drupal (www.drupal.org) and the RDF modules in Drupal [4] to implement eXframe. We developed Drupal "content types" for experiments, biomaterials and assays and mapped these and their attributes to existing biomedical ontologies - primarily the Ontology for Biomedical Investigation (OBI) [5] and the Experimental Factor Ontology (EFO) [6]. The details of the model are described in our paper on eXframe [3]. The biomaterial (sample) attributes were also mapped to ontologies such as cell type to CL, the Cell Type Ontology [7], tissue to FMA, the Foundation Model of Anatomy [8] and disease state to DO, the Disease Ontology [9]. Recently, the model was extended to represent complex samples such as induced pluripotent stem (iPS) cells. The Drupal RDF modules are used to publish Linked Data (RDF), which is indexed by the ARC2 PHP libraries into a SPARQL endpoint. Currently we are collaborating to develop a Fairport (http://datafairport.org) interface for eXframe.

The eXframe platform was successfully used to build the Stem Cell Commons (SCC, http://stemcellcommons.org) repository for genomics data at the Harvard Stem Cell Institute (HSCI) [10]. A screenshot of the published RDF for a sample experiment in SCC is shown in Figure 1.

eXframe is also currently being used to develop similar databases at other institutions. We will present sample SPARQL queries on the Stem Cell Commons SPARQL endpoint that integrates multiple ontologies. The SCC resource is used for a variety of use cases such as i) retrieving experiments performed on a certain cell type in a model organism; ii) displaying assays done on a particular disease model or iii) finding transcript factor binding measurements using next generation sequencing - such use cases will be demonstrated.

## 3 Conclusions

Our reusable platform can be used to build Semantic Web platforms for genomics data allowing flexible queries and integration with other ontologies. The software is freely available at https://github.com/mindinformatics/exframe, under the GPL version 2 free software license.

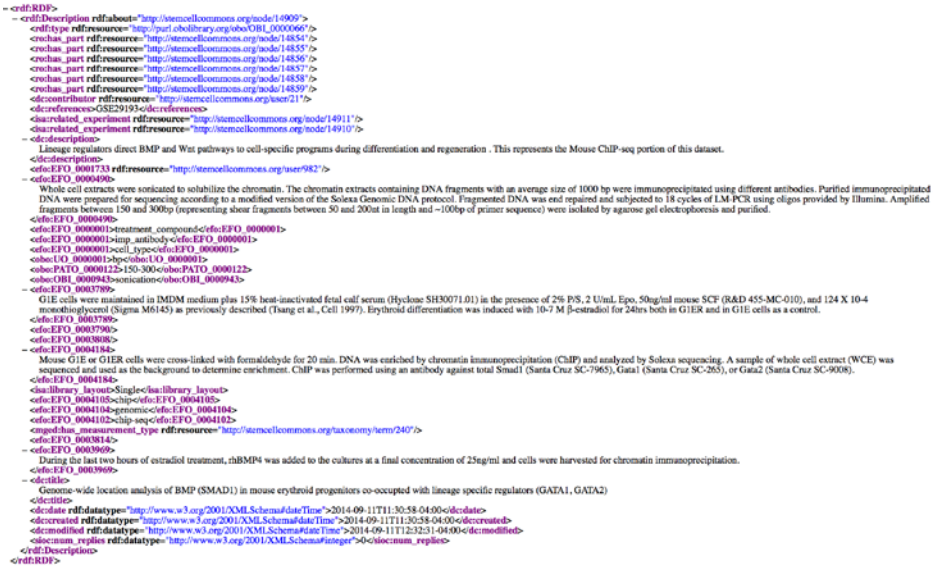


**Figure 1: RDF of a sample experiment**

# 4 References


1. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, et al: Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* 2014, 42:D926-932.
2. Sinha AU, Merrill E, Armstrong SA, Clark TW, Das S. eXframe: reusable framework for storage, analysis and visualization of genomics experiment. BMC Bioinformatics. 2011 Nov 21;12:452. doi: 10.1186/1471-2105-12-452.
3. Merrill E, Corlosquet S, Ciccarese P, Clark T, Das S: Semantic Web repositories for genomics data using the eXframe platform**.** *J Biomed Semantics* 2014, **5:**S3.
4. Corlosquet S, Delbru R, Clark TW, Polleres A, Decker S: Produce and Consume Linked Data with Drupal. 8th International Semantic Web Conference (ISWCC) Washington DC 2009.
5. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, et al: Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 2010, 1 Suppl 1:S7.
6. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H: Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 2010, 26:1112-1118.
7. Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, Mungall CJ, Diehl AD: Logical development of the cell ontology. BMC Bioinformatics 2011, 12:6.
8. Golbreich C, Grosjean J, Darmoni SJ: The Foundational Model of Anatomy in OWL 2 and its use. Artif Intell Med 2013, 57:119-132.

9. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA: Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 2012, 40:D940-946.
10. Ho Sui S, Merrill E, Gehlenborg N, Haseley P, Sytchev I, Park R, Rocca-Serra P, Corlosquet S, Gonzalez-Beltran A, Maguire E, Hofmann O, Park P, Das S, Sansone SA, Hide W. The Stem Cell Commons: an exemplar for data integration in the biomedical domain driven by the ISA framework. AMIA Jt Summits Transl Sci Proc. 2013 Mar 18;2013:70.