Semantic Web 0 (0) 1
IOS Press

Translational research combining orthologous genes and human diseases with the OGOLOD dataset

Editor(s): Pascal Hitzler, Kno.e.sis Center, Wright State University, USA; Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited review(s): Sören Auer, University of Leipzig, Germany; Erik Wilde, UC Berkeley, USA; one anonymous reviewer

José Antonio Miñarro-Giménez [a,*], Mikel Egaña Aranguren [b,d] Boris Villazón-Terrazas [c,d]
Jesualdo Tomás Fernández Breis [a]

[a] *School of Computer Science, University of Murcia (UM), Spain*
[b] *Biological Informatics Group, Centre for Plant Biotechnology and Genomics (CBGP), Technical University of Madrid (UPM), Spain*
[c] *iSOCO, Avda. Partenón 16-18, 28042, Madrid, Spain*
[d] *Ontology Engineering Group (OEG), School of Computer Science, Technical University of Madrid (UPM), Spain*

Abstract. OGOLOD is a Linked Open Data dataset derived from different biomedical resources by an automated pipeline, using a tailored ontology as a scaffold. The key contribution of OGOLOD is that it links, in new RDF triples, genetic human diseases and orthologous genes, paving the way for a more efficient translational biomedical research exploiting the Linked Open Data cloud.

Keywords: Linked Open Data, Linked Data, orthologous genes, human genetic disease, translational research, biomedicine, RDF, SPARQL, OWL, ontology

1. Introduction

OGOLOD is a Linked Open Data (LOD) dataset that represents combined information about orthologous genes and human genetic diseases (Table 1). Orthologous genes present similarity due to the fact that they descend from a common ancestor [4]. Such genes can be used to extrapolate information from one organism to another; hence the interest of using them to infer new hypotheses about human genetic diseases, that is, understanding the genetic causes of diseases [1,6]. The growing interest for facilitating an efficient access to information about orthologous genes is also demonstrated by efforts like Quest for Orthologs [4].

OGOLOD is the output of an automated pipeline that obtains and adapts information from the following sources:

- Databases of orthologous genes: COG [15], Inparanoid [11], Homologene[1], and OrthoMCL [3].
- OMIM[2] (Online Mendelian Inheritance in Man), a database holding information about human genetic diseases.

The information from these sources is collected as RDF instances using the OGO ontology (Figure 1). Thus, the OGO ontology provides the backbone in which to embed all the data collected from the sources: an axiomatised representation of the knowledge do-

*Corresponding author. E-mail: jose.minya@gmail.com.

[1] http://www.ncbi.nlm.nih.gov/homologene
[2] http://www.ncbi.nlm.nih.gov/omim/

main of orthologous genes and human genetic diseases. The OGO ontology combines widely used vocabularies from the life sciences domain, in doing so enhancing the capability of the OGOLOD dataset of being integrated with the rest of the (life sciences) LOD cloud. Such reused vocabularies describe different domains of interest: GO (gene product function, location, or process) [5], ECO[3] (evidence code types that support the link of a gene product with a GO term), NCBI taxonomy (classification of organisms) [13], RO (a set of widely used biological relations) [14], and HPO (a vocabulary for human phenotypic abnormalities) [12]. OGOLOD has also links to other LOD datasets, providing a unique entry point to combine information about orthologous genes with other extant LOD datasets.

OGOLOD allows researchers to combine, in their research hypotheses, information about the evolutionary relationship of genes, their biological implications and their relation with disorders. For instance, researchers of genetic disorders could find genes in other species with similar function to the ones associated with such disorders. As a consequence of such integration of information, OGOLOD gives the researchers the ability to answer queries like "diseases in which the orthologous genes of gene X are involved" (Figure 3), or "all orthologous genes that belong to *Rattus Norvegicus* and are related to the genes involved in lung cancer" (Figure 4).

OGOLOD is the LOD version of the OGO Knowledge Base (KB) [8,9] ("OGOLOD" stands for "OGO Linked Open Data"). OGOLOD has been produced by following the methodological guidelines described in [16] for publishing LOD datasets, committing to an iterative and incremental life cycle model. The process of converting OGO to OGOLOD and the resulting OGOLOD dataset are detailed in [10]; this paper focuses on the OGOLOD dataset alone, providing a self-contained and brief overview.

2. Discussion

2.1. URI design

The URI design in OGOLOD is based on slash URIs, instead of hash URIs, since the dataset has a considerable size. Also, the URIs reflect the structure of the model, making a distinction between classes and instances, to improve the usability of the dataset.[4]

Table 1
Basic information and metrics of the OGOLOD dataset. In the case of outbound links, the `owl:sameAs` relationships were established manually, by adding a relation when the original database ID was the same in both datasets.

Name	**OGOLOD**
Website	`http://miuras.inf.um.es/~ogo/ogolod.html`
SPARQL endpoint	`http://miuras.inf.um.es/sparql`
Datahub entry	`http://thedatahub.org/dataset/ogolod`
RDF dumps	`http://miuras.inf.um.es/dump/ogolod.zip`
Logs	`http://miuras.inf.um.es/sparqlLogs/query.zip`
Sitemap	`http://miuras.inf.um.es/sitemap_index.xml` (It is being indexed at `http://sindice.com/`)
Version (2012/5/10)	0.1
Number of triples	38,035,102
Number of links to `bio2rdf-chebi`	3,456,570
Number of links to `bio2rdf-omim`	18,141
Model expressivity	$\mathcal{ALC}(\mathcal{D})$
License	CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

2.2. Reification of n-ary relations to avoid blank nodes

Blank nodes, *i.e.*, resources without a URI, should be avoided when publishing LOD information [7], which poses a problem when N-ary relations must be represented in the published dataset: the most straightforward way of representing N-ary relations is by using blank nodes.[5] This is the case for the OGOLOD dataset, since the relation between a gene and its GO term (location, function, or process) is qualified with an evidence code (*e.g.*, a gene participates in a concrete process, and such relation is backed by an experimental confirmation). This problem was solved in OGOLOD by merging the OWL ob-

[3] `http://www.obofoundry.org/cgi-bin/detail.cgi?id=evidencecode`

[4] *e.g.*, the URI for the class `Gene` is `http://miuras.inf.um.es/ogolod/ontology/Gene`, and the URI for the Gene `67440` is `http://miuras.inf.um.es/ogolod/resource/Gene/67440`: note the distinction between `ontology` and `resource`.

[5] `http://www.w3.org/TR/swbp-n-aryRelations/#anonvnamed`

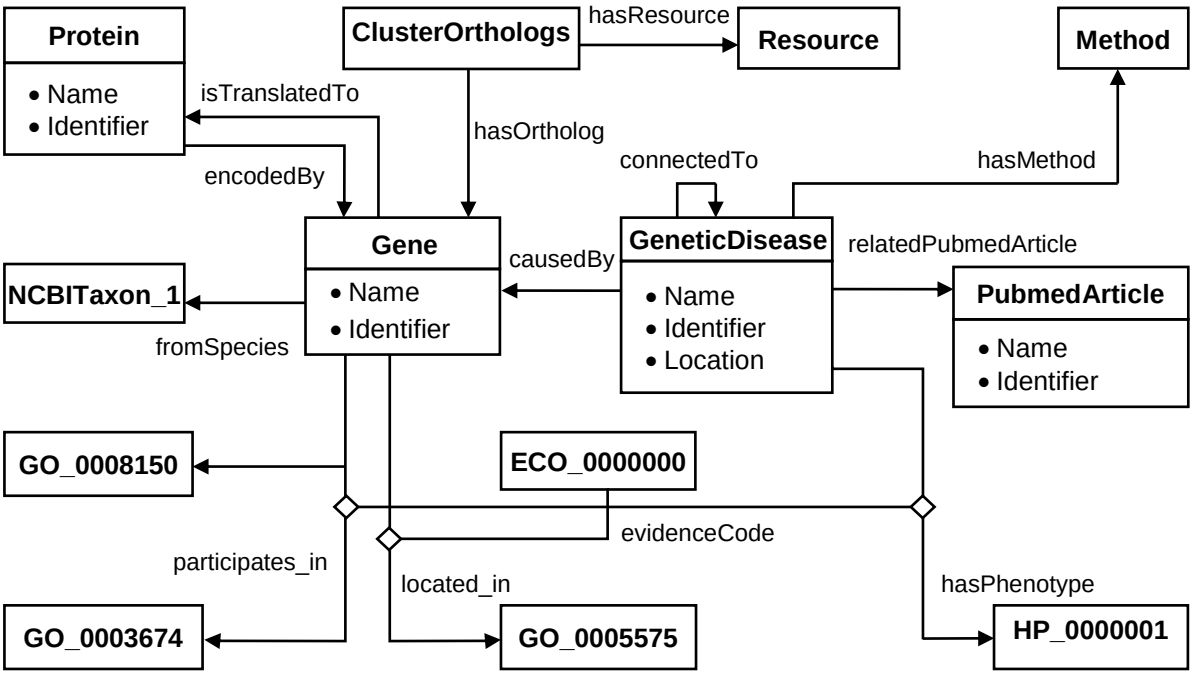

Fig. 1. OGO ontology. Figure reproduced from [10].

ject properties with the evidence codes, *i.e.*, expanding the properties according to the evidence codes: a subproperty of `participates_in` was generated for every evidence code, *e.g.*, `participates_EXP_in` for `Inferred from experiment`, `participates_IMP_in` for `Inferred from mutant phenotype`, and so on. The same procedure was applied in the case of `located_in`.

2.3. Refactoring OWL Punning

The original OGO KB exploited OWL punning[6] in order to translate ontologies written in the OBO format[7] to OWL. In the OGO KB each OBO term is represented as an entity with a URI, and depending on the particular use it can be considered as a class or as an instance. In other words, the same GO term is seen as a class when consulting the class hierarchy whereas it is seen as an instance when it is related to a particular gene through the relationships `participates_in` or `located_in`.

Therefore, the original OBO semantics is preserved but we are able to perform succinct SPARQL queries (without having to represent all the OWL axioms as RDF triples)[8].

When using punning, entities can play different roles and they are differentiated at inference time by their axiomatic context. However, in a LOD setting this is not possible so any entity that exploits punning must be divided in the dataset into different entities (with different URIs) to represent every role. The entities were refactored by the process shown in Figure 2.

2.4. Linking to external datasets

The binding process to link entities of the OGOLOD dataset to external datasets was manually defined. External datasets, such as `bio2rdf`, were examined to identify equivalent entities. The common properties and annotations between internal and external entities were manually identified to guide the binding process. Once the manual examination of the external datasets was completed, custom applications were developed to run this process. A more detailed description of the generation of the OGOLOD dataset is provided in [10].

[6] `http://www.w3.org/TR/owl2-new-features/#F12:_Punning`

[7] `http://www.geneontology.org/GO.format.obo-1_2.shtml`

[8] The general process of translating OBO ontologies to OWL using punning is described in detail in [2].

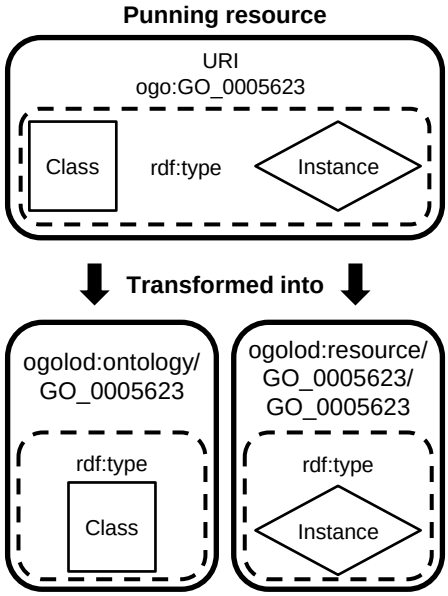

Fig. 2. Refactorisation of URIs of entities that exploit punning. Figure reproduced from [10].

2.5. Known shortcomings of the dataset

The OGOLOD dataset is mainly composed of resources describing clusters of orthologous genes and genetic diseases, which were obtained from the OGO KB. The links with external datasets were added during the translation process, so OGOLOD improves the OGO KB by increasing the number of external links to published datasets.

The OGOLOD dataset has injected new biomedical information that was not available to the LOD cloud. Consequently, the content of the OGOLOD dataset cannot be linked to the original source repositories, but to other datasets that publish resources describing genes and genetic diseases.

Life sciences is one of the most active disciplines in the publication of LOD datasets [7], thus there are many potentially linkable resources. However, the maintenance of high quality links to external datasets is a complex task due to the large size and diversity of such datasets. Therefore, methods to improve the management of such links are needed.

2.6. Use cases

The OGOLOD dataset relates orthologous genes from different species with human genetic disease descriptions. Orthology information is useful to produce predictions of gene function. Hence, obtaining disease information *via* clusters of orthologous genes can be used to find tentative gene functions.

Fig. 3. SPARQL query that relates genetic diseases and genes from *bio2rdf.org*.

```
PREFIX ogolod: <http://miuras.inf.um.es/ogolod/ontology/>
SELECT DISTINCT ?bio2rdf_omim ?title
WHERE {
    ?gene owl:sameAs
       <http://bio2rdf.org/page/geneid:12189> .
    ?cluster ogolod:hasOrtholog ?gene .
    ?cluster ogolod:hasOrtholog ?ortholog .
    ?disease ogolod:causedBy ?ortholog .
    ?disease owl:sameAs ?bio2rdf_omim .
    ?disease ogolod:Name ?title .
}
```

Table 2
Result of the SPARQL query from Figure 3.

?bio2rdf_omim	`http://bio2rdf.org/page/omim:113705`
?title	BREAST CANCER 1 GENE; BRCA1

As an example we propose the search for the human genetic diseases related to the gene `Brca1` of the *Mus musculus* species stored in `bio2rdf.org` datasets. This search is based on the content of OGOLOD and its external links associated with `bio2rdf.org`. Figure 3 shows the SPARQL query associated to this example: in order to define the query the URL of the gene[9] is required as input. To reference the resources from `bio2rdf.org` we exploit the `owl:sameAs` relationships that relate OGOLOD resources with them.

Table 2 shows the result of the query from Figure 3. The resource from `bio2rdf.org`[10] indicates the OMIM resource with identifier `113705` and it corresponds to breast cancer. Thus, we can link the `Brca1` gene from *Mus musculus* species with the Breast cancer disease resource, stored in separate `bio2rdf.org` datasets.

Figure 4 provides another example in which orthologs related to genes involved in lung cancer and belonging to *Rattus Norvegicus* are retrieved. The result can be seen in Table 3.

[9] `http://bio2rdf.org/page/geneid:12189`
[10] `http://bio2rdf.org/page/omim:113705`

Fig. 4. SPARQL query that provides orthologous genes that belong to *Rattus Norvegicus* and are related to the genes involved in lung cancer.

```
PREFIX ogolod:<http://miuras.inf.um.es/ogolod/ontology/>
PREFIX ogolodr:<http://miuras.inf.um.es/ogolod/resource/>
SELECT DISTINCT ?gene2
WHERE {
    ?ortholog ogolod:fromSpecies
        <http://miuras.inf.um.es/ogolod/resource/
        NCBITaxon_10116/NCBITaxon_10116> .
    ?ortholog owl:sameAs ?gene2 .
    ?cluster ogolod:hasOrtholog ?gene .
    ?cluster ogolod:hasOrtholog ?ortholog .
    ?disease ogolod:causedBy ?gene .
    ?disease owl:sameAs
        <http://bio2rdf.org/page/omim:211980> .
}
```

Table 3
Result of the SPARQL query from Figure 4.

gene2
http://bio2rdf.org/page/geneid:363140

3. Conclusion

The OGOLOD dataset integrates information of orthologous genes and human genetic diseases in a LOD setting, making it computer-friendly and discoverable *via* other LOD links. OGOLOD can be used, for example, by scientists interested in studying diseases from a genetic perspective, using orthology information to infer new research paths for such disease. Providing the possibility of performing such analyses is OGOLOD's contribution to the LOD cloud, and more importantly, to translational biomedical research.

Acknowledgements

This work is funded by the Spanish Ministry for Economy and Competitiveness through grant TSI2007-66575-C02-02 and by the EC through the FEDER program. Mikel Egaña Aranguren is funded by the Marie Curie-COFUND Programme (FP7) of the EU. Boris Villazón-Terrazas is funded by the Web n+1 (TSI-020301-2009- 24) and BabelData (TIN2010-17550) projects.

References

[1] E. Antezana, M. Egaña, W. Blondé, A. Illarramendi, I. Bilbao, B. De Baets, R. Stevens, V. Mironov, and M. Kuiper. The Cell Cycle Ontology: An application ontology for the representation and integrated analysis of the cell cycle process. *Genome Biology*, 10(5):R58, 2009.

[2] M. E. Aranguren, J. T. F. Breis, E. Antezana, C. Mungall, A. R. González, and M. Wilkinson. OPPL-Galaxy, a Galaxy tool for enhancing ontology exploitation as part of bioinformatics workflows. *Journal of Biomedical Semantics*, 4(1):2+, 2013.

[3] F. Chen, A. Mackey, C. Stoeckert, and D. Roos. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34(Suppl 1):D363–D368, 2006.

[4] C. Dessimoz, T. Gabaldón, D. S. Roos, E. L. L. Sonnhammer, J. Herrero, and the Quest for Orthologs Consortium. Toward community standards in the quest for orthologs. *Bioinformatics*, 28(6):900–904, 2012.

[5] Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 23(May):25–29, 2000.

[6] K. Harvey, C. Pfleger, and I. Hariharan. The Drosophila Mst ortholog, hippo, restricts growth and cell proliferation and promotes apoptosis. *Cell*, 114:457–467, August 2003.

[7] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011.

[8] J. A. Miñarro Gimenez, M. Egaña Aranguren, R. Martínez Béjar, J. T. Fernández-Breis, and M. Madrid. Semantic integration of information about orthologs and diseases: The OGO system. *Journal of biomedical informatics*, (44):1020–1031, 2011.

[9] J. A. Miñarro Gimenez, M. Madrid, and J. T. Fernandez-Breis. OGO: an ontological approach for integrating knowledge about orthology. *BMC bioinformatics*, 10 Suppl 10(Suppl 10):S13+, 2009.

[10] J. A. Miñarro-Gimenez, M. E. Aranguren, B. V. Terrazas, and J. T. Fernández-Breis. Publishing Orthology and Diseases Information in the Linked Open Data cloud. *Current Bioinformatics*, 7(3):255–266, 2012.

[11] M. Remm, C. V. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052, 2001.

[12] P. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83:610–615, 2008.

[13] E. Sayers, T. Barrett, D. Benson, E. Bolton, S. Stephen, K. Canese, V. Chetvernin, D. Church, M. Deanna, M. DiCuccio, S. Federhen, M. Feolo, L. Y. L.Y.Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. John Wilbur, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 38(suppl 1):D5–D16, 2010.

[14] B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, and C. Rosse. Relations in Biomedical Ontologies. *Genome Biology*, 6:R46,

2005.

[15] R. Tatusov, N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, S. Rao, S. Smirnov, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin, and D. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41–55, 2003.

[16] B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data Linking Government Data. In D. Wood, editor, *Linking Government Data*, chapter 2, pages 27–49. Springer New York, 2011.