

# Publishing Orthology and Diseases Information in the Linked Open Data Cloud

Jose A. Miñarro-Giménez<sup>1</sup>, Mikel Egaña-Aranguren<sup>\*2</sup>, Boris Villazón-Terrazas<sup>2</sup> and Jesualdo T. Fernández-Breis<sup>1</sup>

<sup>1</sup>Facultad de Informática, Universidad de Murcia, Murcia, Spain; <sup>2</sup>Ontology Engineering Group, School of Computer Science, Universidad Politécnica de Madrid, Boadilla del Monte, Spain

**Abstract:** The Linked Data initiative offers a straight method to publish structured data in the World Wide Web and link it to other data, resulting in a world wide network of semantically codified data known as the Linked Open Data cloud. The size of the Linked Open Data cloud, i.e. the amount of data published using Linked Data principles, is growing exponentially, including life sciences data. However, key information for biological research is still missing in the Linked Open Data cloud. For example, the relation between orthologs genes and genetic diseases is absent, even though such information can be used for hypothesis generation regarding human diseases. The OGOLOD system, an extension of the OGO Knowledge Base, publishes orthologs/diseases information using Linked Data. This gives the scientists the ability to query the structured information in connection with other Linked Data and to discover new information related to orthologs and human diseases in the cloud.

**Keywords:** Genetic disease, linked (open) data, orthologs, RDF, semantic web SPARQL and OWL.

## 1. INTRODUCTION

The Semantic Web is a vision for the next generation World Wide Web (WWW) in which structured information will be directly published, instead of because codified in computationally opaque text formats like HTML documents. In a prospective Semantic Web the integration and processing of such information, including automated reasoning would be seamlessly performed by automatic agents [1]. The technology to build the Semantic Web offers a solution to the perennial problem of the life sciences: managing an exponentially growing data deluge. Hence, endeavors like the W3C Semantic Web Health Care and Life Sciences (HCLS) Interest Group [2] offer examples of working semantic technology for life sciences.

The implantation of the Semantic Web in the WWW is a gradual process that relies on a stack of technologies known as the Semantic Web Stack (Fig. 1): in such stack, each technology builds upon the technology below, adding new functionality. Three languages stand out in the stack, from the point of view of the Semantic Web, considering their user base and tool support: RDF [3], SPARQL [4], and OWL [5].

RDF is used to represent data of all sorts using the Subject-Predicate-Object pattern, that is, linking entities in triples (Figs. 2 and 3). SPARQL can be used to query data codified with RDF. OWL is used to codify knowledge (General properties) about the data published with RDF, using ontologies: OWL defines classes (Sets) of entities that share certain features, allowing the use of automated reasoning. OWL sits on top of RDF in the stack;

Both RDF and OWL use URIs[6] to identify entities, to be able to refer to them in a WWW setting.

The most successful step towards the Semantic Web is Linked Data (LD) [7]. LD is a method to publish data on the WWW based on four principles [8]:

1. Use URIs to name entities.
2. Use HTTP URIs so that those names can be resolved.
3. When a user looks up a URI, provide useful information, using RDF and SPARQL.
4. Include links to other URIs to increase the chance of discovering new related data.

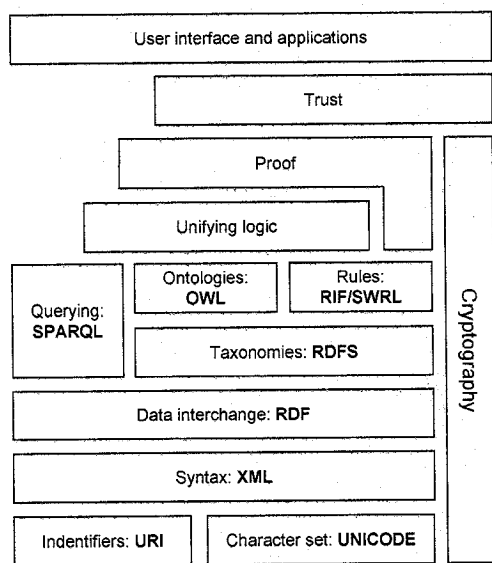
Following such principles any information producer can publish data in RDF and link it to other RDF data. Adding new RDF data to such network is a simple task that follows the same basic principle that made the WWW so successful: new information can be linked to the network without the approval of a central system. Therefore LD exploits the WWW in its current form to offer a single data space in which data can be integrated in a "Pay as you go" fashion [9], i.e. LD works as a "World Wide database".

The LD infrastructure is currently used to link world-wide open data in the Linked Open Data (LOD) cloud (Fig. 4). The data published in the LOD cloud is diverse in granularity, scope, scale and origin: data from governments, geospatial data, media, user-generated data, etc. The LOD cloud is constantly growing and information from new domains is added to it.

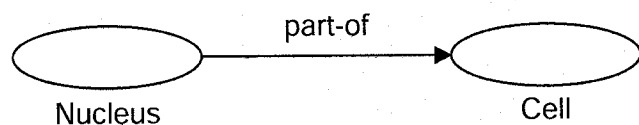
The life sciences community is heavily involved in the development of the LOD cloud: a considerable part of it pertains to life sciences data [11]. This is a logical situation since the main problem of life sciences is the fact that information from many resources needs to be integrated in

\*Address correspondence to this author at the Ontology Engineering Group, School of Computer Science, Universidad Politécnica de Madrid, Boadilla del Monte, Spain; Tel: 0034-656724785; E-mail: megana@fi.upm.es

order to be able to perform research [12]. The LOD cloud offers the ideal meaningful infrastructure for such goal. Additionally, the availability of consensus ontologies generated by the life sciences community, like the ontologies collected in the OBO Foundry [13] (See Section 0), facilitates the publication of data in the LOD cloud, since OWL ontologies are used as vocabularies to refer to when generating RDF data.



**Fig. (1).** The Semantic Web stack [10]. In the stack, each technology adds new functionality, retaining the functionality of the technologies below. For example, OWL can be serialized in RDF/XML syntax, and treated like RDF, with no semantics: tools built to deal with RDF (e.g. SPARQL) are able to process OWL, albeit without exploiting its full expressivity. Both OWL and RDF use URIs to identify entities.

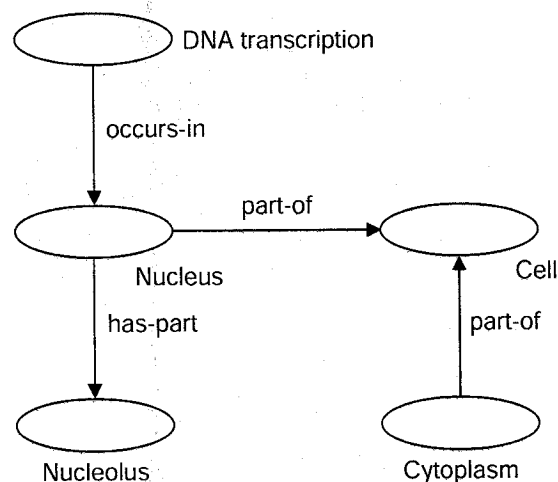


**Fig. (2).** RDF triple, stating that a nucleus is a part of a cell. Subject: nucleus; predicate: part-of; object: cell. The subject, predicate and object are identified by URIs (Not shown) and each of them may belong to a different resource, allowing the linking of the data.

LD offers the possibility of performing queries against the information published, rather than performing keyword based searches, increasing the precision of the retrieved information. Additionally, the navigation through LD links is a much richer experience compared to the navigation through hypertext links, since LD links are semantic, with a precise meaning. Therefore, the information in LD is connected more thoroughly.

Currently, even though life sciences have contributed considerable amounts of data to the LOD cloud, much biological information still remains outside, scattered in different and unrelated online databases. Information relating genetic diseases and orthologs genes, for example, is absent from the LOD cloud. Ortholog genes are genes that belong

to species that diverged from the same ancestor during evolution. Since orthology information can be used to infer further putative information related to a given disease, it is important to have orthology information in the LOD cloud, especially if it is directly related to human diseases.



**Fig. (3).** RDF graph, obtained combining different triples.

Orthology information can be used, for example, to obtain information about orthologs of the genes that cause breast cancer, but such query is not possible in the current LOD cloud. Therefore, if a biomedical researcher is interested in retrieving such information, she will need to use independent tools and resources and to manually combine the information. If all this information would be available in the LOD cloud, this researcher would be able to access such information in a semantically connected and integrated way. Hence, we have built OGOLOD, a LOD dataset about orthologs and diseases. OGOLOD is an extension of the OGO system, a Knowledge Base (KB) that collects orthologs/diseases information and stores it semantically [14].

This paper discusses the content and design of OGOLOD: Section 2 describes the original OGO KB; Section 3 describes the OGOLOD dataset in detail, including the links to other datasets and the decisions made to adapt the OGO KB to a LD setting; Section 4 describes the interfaces that can be used to access the dataset, explaining how, for example, genetic information related to breast cancer can be obtained; Section 5 discusses the implementation of the OGOLOD system, including the tools used; Section 6 provides the benefits of the OGOLOD system, taking into account the current LOD datasets and LD best practices.

## 2. THE OGO SYSTEM

The OGO system was developed to provide the community with an integrated resource for accessing information relating orthologs and genetic diseases. This system allows to associate the genes involved in a particular genetic disorder with their corresponding clusters of orthologs, by integrating the data from the following resources:

Publishing Orthology and Diseases Information

- KOG [16]: The euKaryotic Orthologous Groups of proteins database (KOG) is part of the Cluster of Orthologous Groups (COG) developed by the National Center for Biotechnology Information (NCBI) [17].
- Inparanoid [18]: This resource stores information about clusters of orthologous genes and proteins of 35 organisms.
- Homologene [19]: This database contains information about clusters of orthologous genes.
- OrthoMCL [20]: It contains information about orthologous genes from 87 organisms, classified in groups of orthologous proteins.
- OMIM (Online Mendelian Inheritance in Man) [21]: This NCBI database contains information about human genetic disorders.

The knowledge stored in the OGO system is organized around the OGO ontology, which provides a formal representation of the domain of orthology (orthologs clusters from different databases) and genetic diseases (OMIM records), also defining the schema for the OGO KB. Both the ontology and the KB are described next.

2.1. The OGO Ontology

The OGO ontology was defined to conceptualize the domain of clusters of orthologs and genetic diseases (Fig. 5).

The OGO ontology relates both domains through the concept *Gene*.

The main concepts related to the domain of orthologs are *Gene*, *Protein*, *Cluster Orthologs*, *Resource* and *NCBITaxon\_1*. The gene concept is related to *NCBITaxon\_1*, which represents the species associated with a gene, through the property *from Species*; the protein concept is associated with the gene concept through the properties *is Translated To* and *encoded By*, depending on the direction of the relation. The cluster of orthologs concept is related to the gene concept through the *has Ortholog* property. Hence, the instances of genes that are related to the same instance of cluster of orthologs are considered orthologous genes. The property *has Resource* is used to identify which is the source repository of the orthology information, connecting a cluster of orthologs instance with an instance of the *Resource* concept.

The main concepts related to the domain of genetic diseases are *Genetic Disease*, *Gene*, *Method* and *Pubmed Article*. The *Genetic Disease* concept represents the records of genetic diseases as defined in OMIM. We use the *caused by* property to associate an instance of genetic disease with the instance of gene whose abnormalities are considered a cause of the disease. The *Method* concept represents the diverse research methods to investigate the origin of the genetic diseases, such as "Deductions from the amino acid

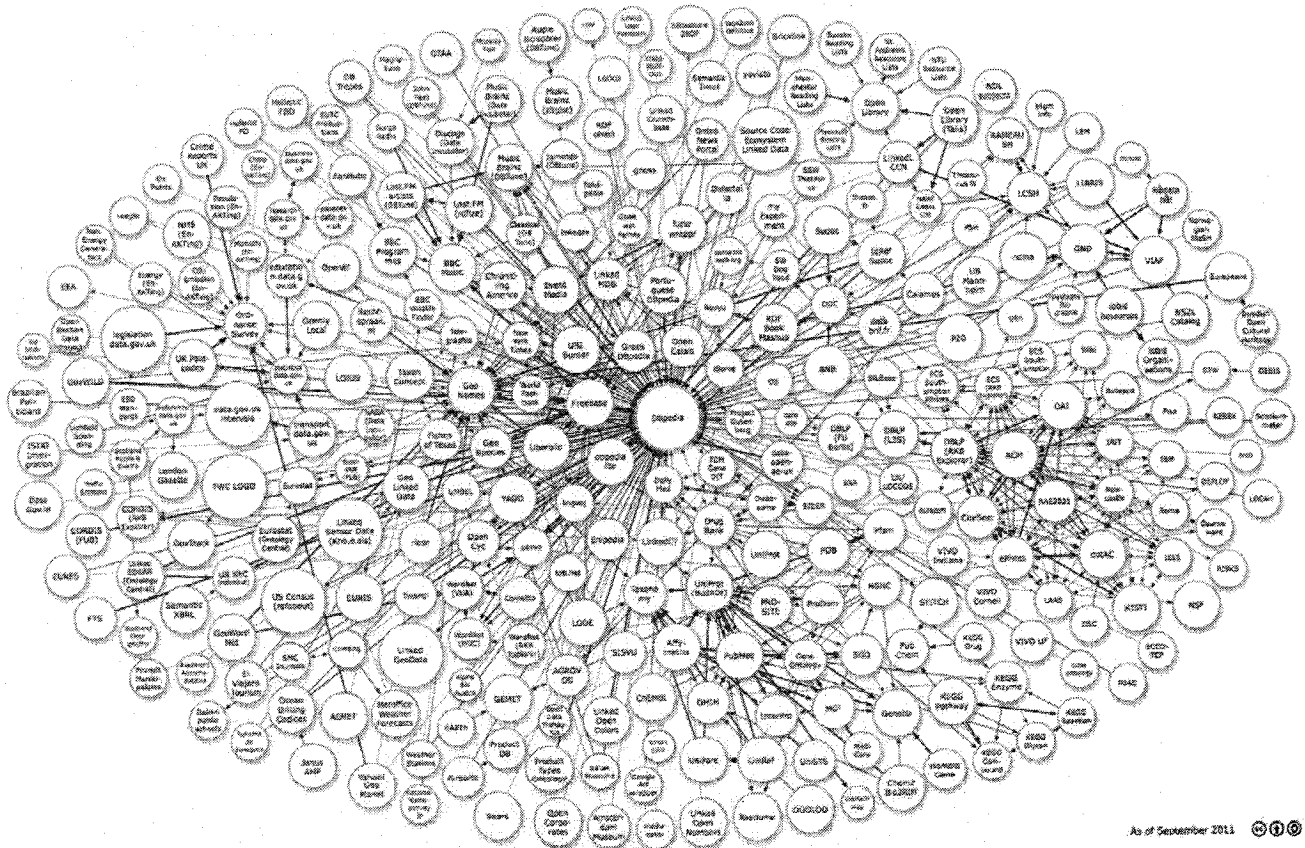


Fig. (4). The LOD cloud [15]. The DBpedia (A LD version of Wikipedia) is the central hub of the LOD cloud. Each bubble is a dataset, its size being proportional to the number of triples. The arrows represent multiple predicates. The OGOLOD dataset, presented in this paper is highlighted by the thick arrow.

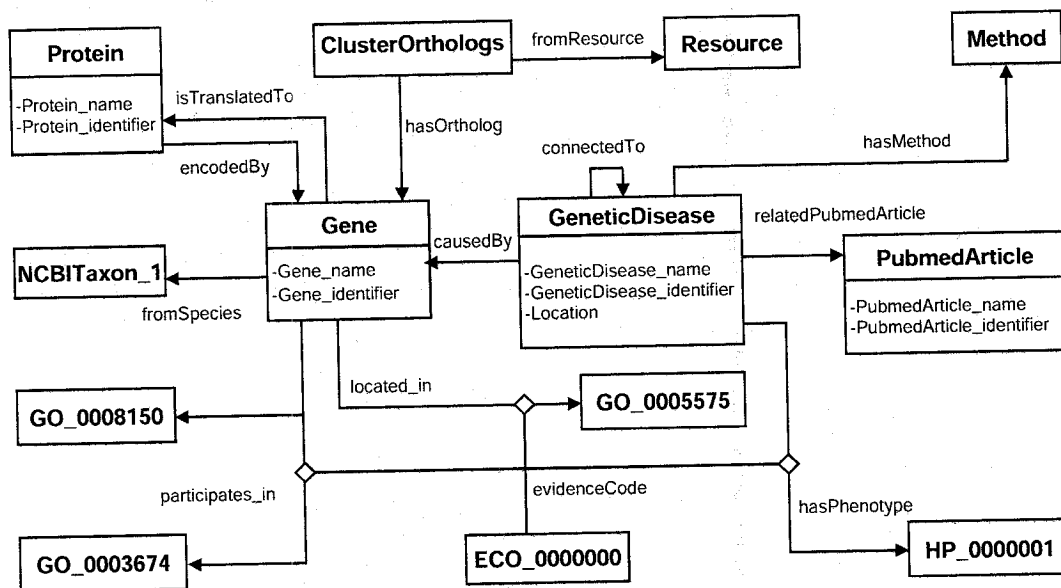


Fig. (5). The OGO ontology provides a formalization of the orthologs/genetic disorders domain in order to integrate such information.

sequence of proteins". Each genetic disease record is associated with a set of biomedical papers citations stored in PubMed [22]. This connection is made through the property *related Pubmed Article*.

The OGO ontology was additionally enriched by reusing the following bio-ontologies to build it, as recommended by current best practices in bio-ontology engineering [23]:

**Gene Ontology (GO):** GO provides a structured controlled vocabulary to describe the cellular component (GO\_0005575), molecular function (GO\_0003674) and biological process (GO\_0008150) associated with gene products.

**Evidence Codes Ontology (ECO):** Gene Ontology Annotations (GOA) link gene products of different databases with GO terms [24]. Such links are qualified with evidence codes that define how such relations were obtained (e.g. Inferred from Physical Interaction, Inferred from Genomic Context, etc.). ECO [25] provides a hierarchy of evidence codes.

**NCBI taxonomy database (NCBI):** The NCBI taxonomy [26] represents the taxonomical classification of biological species.

**Relationship Ontology (RO):** RO [27] was created in order to provide a common set of relationships for bio-ontologies, to facilitate their integration. We used the RO *located\_in* and *participates\_in* relationships to associate GO terms with gene products.

**Human Phenotype Ontology (HPO):** HPO [28] provides a standardized vocabulary of phenotypic abnormalities. The addition of HPO is the newest extension of the OGO ontology and it is related to the genetic diseases instances through the *hasPhenotype* relationship.

## 2.2. The OGO Knowledge Base

The OGO KB was obtained as a result of applying the integration methodology defined in [14] to the available information relating orthologs and diseases. Such methodo-

logy provides guidelines to integrate independent and/or overlapped biomedical information resources, based on defining the mappings between the repository schemas of the biomedical resources and the domain ontology. Later on, an automatic integration of the information in the KB is performed, codifying the information as ontology instances and avoiding redundancy and incoherence. Therefore, the resulting OGO KB exploits its ontology model to represent information precisely. Additionally, the information is semantically annotated, so other applications can query and use it.

The current OGO KB contains more than 50,000 orthologs clusters, more than a million genes and proteins, roughly 18,000 human genetic disorders and more than 100,000 references to PubMed citations. The users can exploit this KB through two methods: users may input a series of keywords [29] or define ontology-driven semantic queries [30], [31]. Besides, external applications may consult the KB deploying Web Services [32].

## 3. THE OGOLOD DATASET

In this section we describe how the original OGO system was re-engineered to publish its content in the LOD cloud. The result of this process is an RDF dataset describing the information stored in the OGO system, ready to be published in the LOD cloud, named OGOLOD.

In order to publish the OGOLOD dataset in the LOD cloud we applied the LD principles to the OGO KB. Thus, we designed and implemented a method to analyze and translate the content of the OGO KB into the OGOLOD dataset. The aim of this transformation method is to translate an ontological model and its instances into RDF triples that are ready to be published in the LOD cloud. The KB provides a formal representation of its domain knowledge, which can be processed and consulted by third party applications. Therefore, applications can gather the instances of a particular concept, such as all instances of the gene concept, and search through the ontology model for its

related instances, such as the GO terms related to a particular gene. Since the OGO KB was defined by using OWL, its transformation into the OGOLOD dataset was less laborious than from other types of repositories like relational databases. However, we had to adapt the content of the OGO KB to fulfill the LD principles, which required performing the tasks described in the following sections: Avoiding blank nodes, Transformation of URIs, Associating OGOLOD instances with LOD instances, and refactoring OWL punning.

### 3.1. Avoiding Blank Nodes

Blank nodes are resources that lack a public URI. Guidelines for publishing information as LD [33] discourage the use of blank nodes because it is impossible to set external RDF links to a blank node, and merging data from different sources becomes difficult. In the case of OGOLOD, the triple links between genes, evidence codes and GO terms were refactored to satisfy this requirement. Gene products are related to GO terms through *participates\_in* and *located\_in* relationships. Furthermore, evidence codes represent the experimental evidence for these relationships: a gene product participates in a biological process with a certain evidence. This structure was represented in the OGO KB using blank nodes, where gene instances were related to blank nodes and the blank nodes were linked to the corresponding GO terms and evidence codes. However, we extended the *participates\_in* and *located\_in* relationships by adding a hierarchy of subproperties that represent each type of evidence code. Thus, *participates\_EXP\_in*, which is a subproperty of *participates\_in*, is used for a connection between a gene, a GO term and the evidence code *EXP* (Inferred from experiment). This solution was also applied to the triple relationship between a genetic disease, a human phenotype and an evidence code.

### 3.2. Transformation of URIs

The URIs of the OGOLOD dataset must be defined to be accessed with the HTTP protocol. The URIs of the resources had to be changed in order to deploy the Web server of the OGOLOD system. Besides, to get more manageable URIs, we added in the definition of the URIs the label "ontology/" or the label "resource/", depending whether the data belongs to the schema or to the instances. For example the URI <http://miuras.inf.um.es/ogolod/ontology/Gene> indicates the class Gene of the ontology and the URI <http://miuras.inf.um.es/ogolod/resource/Gene/67440> indicates the instance of the class Gene with the gene identifier "67440". By adding this information to the URIs the users can understand the type of information they are accessing. Moreover, we have included the class of the instances in its URIs to identify the type of the instances. We also replaced some characters with special meaning in the HTTP URL specification, such as "#" or ":", to prevent the use of characters with different meaning to the one established in the document RFC3986 [34].

### 3.3. Associating OGOLOD Instances with LOD Instances

To link the OGOLOD dataset resources to other datasets in the LOD cloud we had to look for appropriate instances in such datasets. This task was manually performed by searching in the Data Hub [35] search interface of CKAN for

other repositories with the same instances. We found the datasets which contain instances from Homologene, Entrez Gene, Pubmed, OMIM, GO, NCBI taxonomy, Uniprot KB, Entrez Protein and HPO bioinformatics repositories and bio-ontologies. These external repositories are part of the Bio2RDF [36] and Linked Life Data [37] projects. Table 1 shows an example of the URIs of external datasets and the equivalent URIs of the OGOLOD repository for each integrated bioinformatics repository and bio-ontology.

**Table 1. Links to External Resources from the OGOLOD Dataset. The Top Table Provides the URIs and Prefixes Used in OGOLOD; the Bottom Table Uses Such Prefixes to Describe the Links. The First Three Links (Gene Protein, and caused by) are Vocabulary Level Links; the Rest are Examples of Instance Level links**

Prefix	IRI
ogoont:	< <a href="http://miuras.inf.um.es/ogolod/ontology/">http://miuras.inf.um.es/ogolod/ontology/</a> >
ogores:	< <a href="http://miuras.inf.um.es/ogolod/resource/">http://miuras.inf.um.es/ogolod/resource/</a> >
uniprot:	< <a href="http://purl.uniprot.org/core/">http://purl.uniprot.org/core/</a> >
ro:	< <a href="http://www.obofoundry.org/ro/ro.owl#">http://www.obofoundry.org/ro/ro.owl#</a> >
b2r:	< <a href="http://bio2rdf.org/">http://bio2rdf.org/</a> >
llfd:	< <a href="http://linkedlifedata.com/resource/phenotype/">http://linkedlifedata.com/resource/phenotype/</a> >
OGOLOD URI	External URI
ogoont:Gene	uniprot:Gene
ogoont:Protein	uniprot:Protein
ogoont:causedBy	ro:has agent
ogores:ClusterOrthologs/10008	b2r:homologene:10008
ogores:Gene/67440	b2r:page/geneid:67440
ogores:PubmedArticle/15902763	b2r:pubmed:15902763
ogores:GeneticDisease/100070	b2r:omim:100070
ogores:GO_0032283/GO_0032283	b2r:go:0032283
ogores:Protein/Q2IHD7	b2r:uniprot:Q2IHD7
ogores:NCBITaxon_9606/NCBITaxon_9606	b2r:taxonomy:9606
ogores:Protein/285813127	b2r:protein:285813127
ogores:HP_0000082/HP_0000082	llfd:id/HP:0000082

Since the OGOLOD dataset and the external datasets share the identifiers and accession codes used by the biomedical resources, we can automatically transform the resource URIs of the OGOLOD into the external datasets URIs. Thus, we linked the OGOLOD URIs to the generated URIs of external resources through the *owl:sameAs* relationship for instances, and the *owl:EquivalentClass* and *owl:EquivalentProperty* relationships for ontology classes and properties definitions. Fig. (6) shows examples of the different types of links of the OGOLOD dataset to external datasets. For example, the *ogolod:NCBITaxon\_7227/-NCBITaxon\_7227* instance identifies the NCBI species record of the *Drosophila melanogaster* species with the identifier "7227". Therefore, the OGOLOD URI can be translated into the Bio2RDF dataset URI for the NCBI

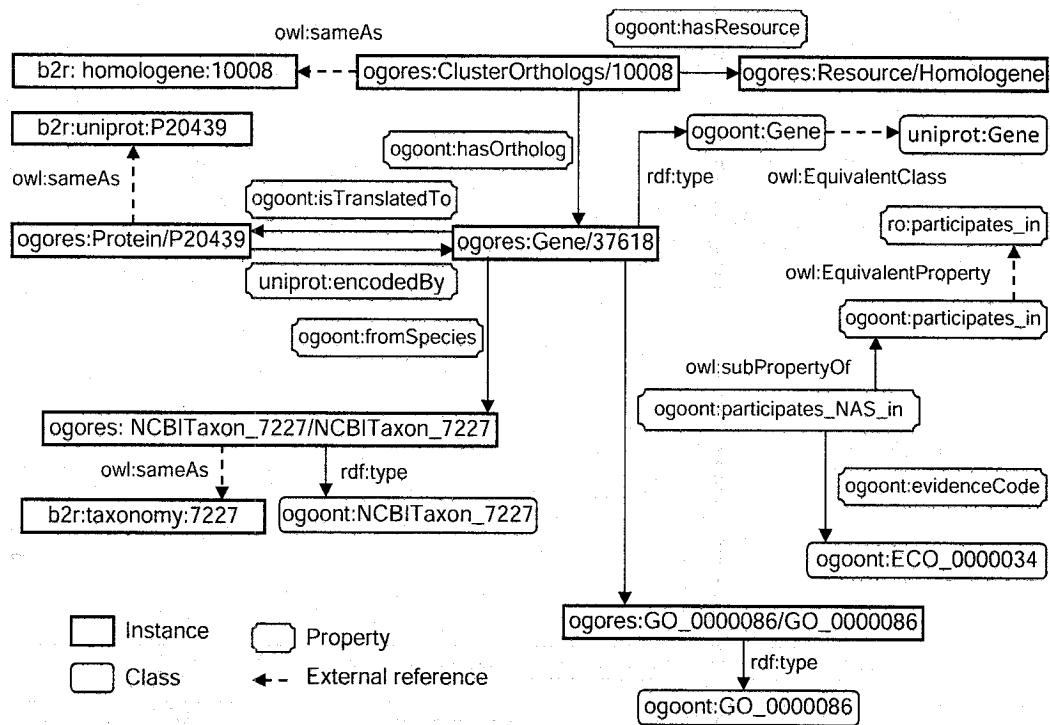


Fig. (6). Excerpt of the OGOLOD dataset. Internal/external resources and vocabulary level/instance level links are shown.

organism classification by using the organism identifier with the suitable Bio2RDF namespace to obtain the *bio2rdf:taxonomy:7227* URI. Finally, we state that both URIs refer to the same instance through the *owl:sameAs* property. The mappings of OGOLOD data with other LOD datasets are obtained by following a similar process.

### 3.4. Refactoring OWL Punning

An important issue we had to face when translating the OGO KB into the OGOLOD dataset was adapting the part of the OGO KB that was modeled by using OWL punning to an LD setting. OWL punning allows using the same URI for different entities [38]: in the case of the OGO KB it was deployed to add an instance to every class with the same URI. Thus, the resource identified by an URI is considered a class or an instance depending on its axiomatic context. Punning was used due to the fact that the imported OBO ontologies only defined hierarchies of classes and the OGO KB required the existence of instances in the model. Therefore, to create the OGOLOD dataset we had to use different URIs for punned instances and classes, because RDF-based data exploitation cannot use the axiomatic context. Thus, we extended the imported bio-ontologies with their corresponding instances for each concept in the OGOLOD dataset. Fig. (7) shows an example of how the punned resource *ogo:GO\_0005623* of the OGO KB, which represents the *GO:0005623* of GO, is transformed into two resources in the OGOLOD dataset: the *ogolod:ontology/-GO\_0005623* class and the *ogolod:resource/GO\_0005623/-GO\_0005623* instance.

## 4. OGOLOD INTERFACES

Apart from the raw RDF dataset, available in CKAN (Comprehensive Knowledge Archive Network) [39],

OGOLOD offers two interfaces to access the data: the SPARQL interface and the navigation interface.

### 4.1. SPARQL Interface

The SPARQL interface [40] can be used to perform SPARQL queries against the OGOLOD dataset. Our running example is the study of breast cancer mentioned in Section 1; we are interested in retrieving information about the genes associated with such disease, and their orthologs. Section 2 shows the SPARQL query that would provide such information; we are searching for the name and identifier of the genes associated with breast cancer. In this query, we have defined both the name and OMIM identifier of the disease, although the same results would be obtained by using only the name or the identifier. The results of this query are the following genes (all of them having the same code, 83990): OF, BRIP1, BACH1, MGC126521, FLJ90232, MGC126523, FANCF, ENSG00000136492.

We are also interested in knowing which genes are orthologous to the ones associated with breast cancer. The appropriate SPARQL query for the gene BRIP1 is shown in Table 3 *gene1* refers to the gene BRIP1 and *?gene2* is any gene orthologous to it. The results of the query also include the URI of the taxon of *?gene2*. A snapshot of the results is shown in Table 4.

We are also interested in getting more information about BRIP1: for instance, the evidence codes for the relationships in which BRIP1 is involved in humans. The corresponding query is shown in Table 5. In such query blank nodes are avoided through the use of the *rdfs:subPropertyOf* property. The results to this query are shown in Table 7.

**Table 2. SPARQL Query for Obtaining the Genes Associated with Breast Cancer**

```
PREFIX ogolod:<http://miuras.inf.um.es/ogolod/ontology/>
SELECT DISTINCT ?geneID ?geneName
WHERE {
  ?disease ogolod:causedBy ?gene .
  ?disease ogolod:Name "BREAST CANCER" .
  ?disease ogolod:Identifier "11480" .
  ?gene ogolod:Identifier ?geneID .
  ?gene ogolod:Name ?geneName
}
```

**Table 3. SPARQL Query for Obtaining the Genes that are Orthologous to BRIP1, which is Associated with Breast Cancer**

```
PREFIX ogolod:<http://miuras.inf.um.es/ogolod/ontology/>
SELECT DISTINCT ?geneID ?geneName ?specie "BRIP1"
WHERE {
  ?cluster_Orthologs ogolod:hasOrtholog ?gene1 .
  ?cluster_Orthologs ogolod:hasOrtholog ?gene2 .
  ?gene1 ogolod:Name "BRIP1" .
  ?gene2 ogolod:Identifier ?geneID .
  ?gene2 ogolod:Name ?geneName .
  ?gene2 ogolod:fromSpecies ?specie
}
```

Finally, we are interested in finding the molecular functions in which BRIP1 is involved. This will allow the prospective biomedical researcher to know the activities in which the gene that causes the disease participates. The corresponding query is shown in Table 6. This example uses the TRANSITIVE option, making it possible to exploit the GO hierarchy. It should be noted that the TRANSITIVE option is "syntactically" implemented by Virtuoso, thus it is not exploiting OWL semantics for inference. The resulting terms are shown in Table 8.

**4.2. Navigation Interface**

The OGOLOD data can be explored through an HTML interface that can be used to navigate to external resources

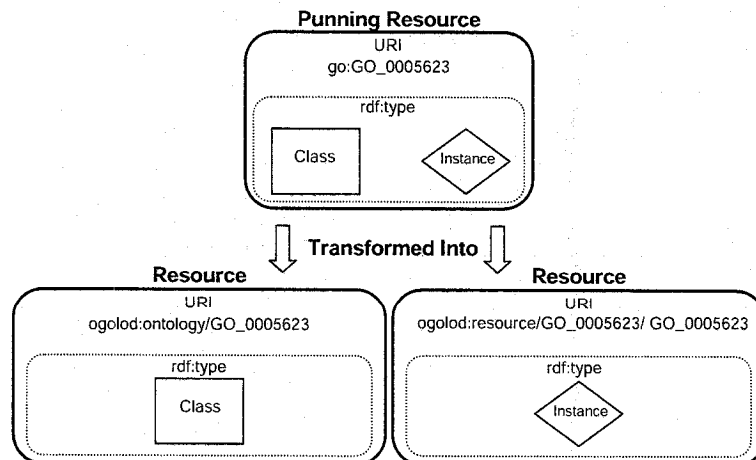
by using RDF links. For instance, Fig. (8) is a partial view of the navigation screen for breast cancer. This interface can be used to navigate through the dataset, via internal links (e.g. *ogolodonto:causedBy*) or external links (e.g. *owl:sameAs*). In Fig. (8) we can see the name of the disease, the chromosomal locations, the genes involves in the disease, and so forth.

**Table 4. Genes that are Orthologous to BRIP1**

geneID	geneName	specie
173370	F33H2.1, dog-1	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_6239/NCBITaxon_6239
237911	OF, Brip1, Bach1, RP23-36C4.3, 3110009N10Rik, 8030460J03Rik, ENSMUSG00000034329	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_10090/NCBITaxon_10090
360588	Brip1, ENSRNOG00000027475	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_10116/NCBITaxon_10116
417642	BRIP1, ENSGALG00000005279	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_9031/NCBITaxon_9031
454796	BRIP1, ENSPTRG00000009493	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_9598/NCBITaxon_9598
491122	BRIP1	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_9615/NCBITaxon_9615
794038	fancj, brip1, si:ch211-15811.1	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_7955/NCBITaxon_7955
83990	OF, BRIP1, BACH1, MGC126521, FLJ90232	http://miuras.inf.um.es/ogolod/resource/NCBITaxon_9606/NCBITaxon_9606

**5. IMPLEMENTATION**

In this section we present the process followed for publishing OGOLOD as LD, which is an adaptation of the LD lifecycle activities proposed in [41]. The lifecycle



**Fig. (7).** Transformation of OGO OWL punning into the OGOLOD dataset.

activities Specification and Publication are described next.

**Table 5. SPARQL Query for Obtaining the Evidence Codes of the Relationships the Human BRIP1 is Involved in**

```
PREFIX ogolod:<http://miuras.inf.um.es/ogolod/ontology/>
SELECT distinct ?o ?g ?l
WHERE {
  ?o ogolod:Name "BRIP1" .
  ?o ogolod:fromSpecies ?speciel .
  ?speciel rdfs:label "9606" .
  ?o ?p ?g .
  ?p rdfs:subPropertyOf ?u .
  ?u ogolod:evidenceCode ?e .
  ?e rdfs:label ?l
}
```

**Table 6. SPARQL Query for Obtaining the GO Molecular Functions a Given Gene is Involved in, Including the Evidence Code**

```
PREFIX ogolod:<http://miuras.inf.um.es/ogolod/ontology/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?gene_ID ?relationship ?GO_Name ?eco_Name
WHERE {
  ?gene ogolod:Name "BRIP1" .
  ?gene ogolod:Identifier ?gene_ID .
  ?gene ?participates_in ?goTerm .
  ?participates_in rdfs:label ?relationship .
  OPTIONAL {
    ?participates_in ogolod:evidenceCode ?eco .
    ?eco rdfs:comment ?eco_Name .
  }
  ?goTerm rdfs:label ?GO_Name .
  ?goTerm rdf:type ?type .
  ?type rdfs:subClassOf ogolod:GO_0003674 OPTION
  (TRANSITIVE) .
```

**5.1. Specification**

One of the main decisions to be taken in the generation of RDF, independently of the system used for performing this

transformation, is the format in which identifiers (URIs) will be generated. URIs are extremely relevant in this process since they are the key for the alignment of heterogeneous resources that come from different data sources.

**Table 8. Molecular Functions Associated with the Relations in which BRIP1 is Involved**

gene_ID	Relationship	GO_Name
417642	IEA	GO:0000166
417642	participates with inferred from electronical annotation evidence in	GO:0000166
417642	IEA	GO:0003677
417642	participates with inferred from electronical annotation evidence in	GO:0003677
417642	IEA	GO:0016818
417642	participates with inferred from electronical annotation evidence in	GO:0016818
417642	IEA	GO:0005524
417642	participates with inferred from electronical annotation evidence in	GO:0005524
417642	IEA	GO:0004386
417642	participates with inferred from electronical annotation evidence in	GO:0004386
417642	IEA	GO:0004003
417642	participates with inferred from electronical annotation evidence in	GO:0004003

We devoted some effort to choose cool URIs [42] for the OGOLOD resources. For concepts and properties available in our ontologies we use the following pattern:

**Table 7. Evidence Codes Associated with the Relations in which BRIP1 is Involved in Humans**

o	g	l
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0005634/GO_0005634/">http://miuras.inf.um.es/ogolod/resource/GO_0005634/GO_0005634/</a>	ECO:0000006
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0005634/GO_0005634/">http://miuras.inf.um.es/ogolod/resource/GO_0005634/GO_0005634/</a>	experimental evidence
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0005737/GO_0005737/">http://miuras.inf.um.es/ogolod/resource/GO_0005737/GO_0005737/</a>	ECO:0000006
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0005737/GO_0005737/">http://miuras.inf.um.es/ogolod/resource/GO_0005737/GO_0005737/</a>	experimental evidence
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0006357/GO_0006357/">http://miuras.inf.um.es/ogolod/resource/GO_0006357/GO_0006357/</a>	ECO:0000006
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0006357/GO_0006357/">http://miuras.inf.um.es/ogolod/resource/GO_0006357/GO_0006357/</a>	experimental evidence
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0005515/GO_0005515/">http://miuras.inf.um.es/ogolod/resource/GO_0005515/GO_0005515/</a>	ECO:0000006
<a href="http://miuras.inf.um.es/ogolod/resource/Gene/83990">http://miuras.inf.um.es/ogolod/resource/Gene/83990</a>	<a href="http://miuras.inf.um.es/ogolod/resource/GO_0005515/GO_0005515/">http://miuras.inf.um.es/ogolod/resource/GO_0005515/GO_0005515/</a>	experimental evidence



at miuras.inf.um.es  
<http://miuras.inf.um.es/ogolod/resource/GeneticDisease/114480>

Property	Value
ogolod:Identifier	■ 114480
ogolod:Location	■ 11p15.5 ■ 11q22.3 ■ 12p12.1 ■ 13q12.3 ■ 14q32.3 ■ 15q15.1 ■ 16p12 ■ 16q22.1 ■ 17p13.1 ■ 17q21 ■ 17q22 ■ 17q22-q23 ■ 1p32 ■ 22q12.1 ■ 2q33 ■ 2q34-q35 ■ 3q26.3 ■ 5q33.2 ■ 6p25 ■ 8q11
ogolod:Name	■ BREAST CANCER
ogolod:causedBy	■ ogores:Gene/83990
ogolod:hasMethod	■ ogores:Method/A ■ ogores:Method/Ch ■ ogores:Method/D ■ ogores:Method/EM ■ ogores:Method/Fd ■ ogores:Method/H ■ ogores:Method/Pcm ■ ogores:Method/Psh ■ ogores:Method/R ■ ogores:Method/REa ■ ogores:Method/REc ■ ogores:Method/REn ■ ogores:Method/S
ogolod:hasPhenotype_EA	■ ogores:HP_0000006/HP_0000006 ■ ogores:HP_0001425/HP_0001425 ■ ogores:HP_0003002/HP_0003002

Fig. (8). HTML interface of OGOLOD for breast cancer.

- <http://miuras.inf.um.es/ogolod/ontology/{ontology element}>

For example, <http://miuras.inf.um.es/ogolod/ontology/GeneticDisease> is the URI of the class of genetic diseases. For the instances we use the following pattern:

- <http://miuras.inf.um.es/ogolod/resource/{resource type}/{resource name}>

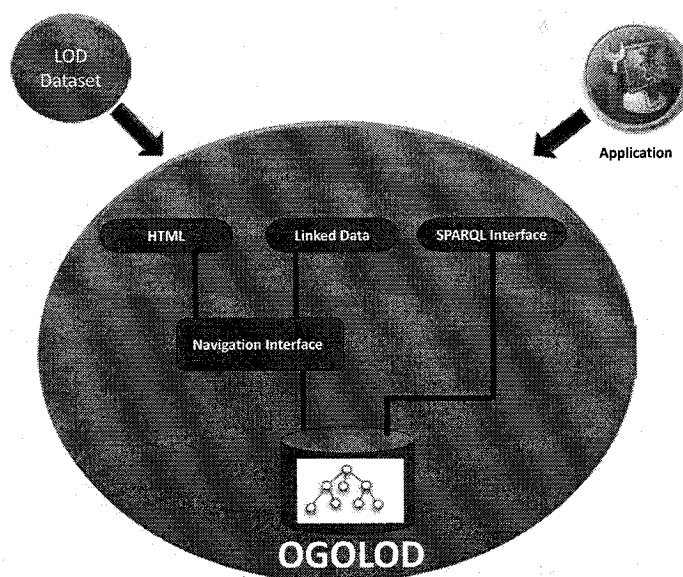
where the resource name is the name of the instance and the resource type is the ontology class name of its type. For example, <http://miuras.inf.um.es/ogolod/resource/GeneticDisease/100070> is the URI of the instance of the genetic disease which is identified by "100070".

It should also be noted that the URI scheme for OGOLOD entities is based on slash URIs instead of hash URIs. In spite of the fact that there is some criticism of the slash URIs because using them requires two HTTP requests to retrieve a single description of an object, slash URIs are appropriate for dealing with resource descriptions that are part of very large datasets [33], as is the case for the OGOLOD dataset.

## 5.2. Publication

Once we have transformed the data to RDF, we need to store and publish that data in a triplestore, *i.e.* a purpose-built database for the storage and retrieval of RDF. There are several tools for storing RDF datasets, for example Virtuoso Universal Server [43], Jena [44], Sesame [45], 4Store [46], YARS [47], and OWLIM [48]. Some of them already include a SPARQL endpoint and LD front-end. However, there are some tools like Pubby [49], Joseki [50], and Talis Platform [51] that provide these functionalities. A thorough overview of the "recipes" for publishing RDF data can be found in [52].

For the publication of the OGOLOD dataset we chose Virtuoso Universal Server, which combines the functionalities of traditional DBMS, virtual databases, RDF triple stores, XML stores, web application servers and file servers. On top of it, Pubby is used for the visualization and navigation of the raw RDF data. Fig. (9) shows the OGOLOD publishing architecture. The last task is to enable the effective discovery and synchronization of our dataset. This is done through the following phases:



**Fig. (9).** Architecture of the OGOLOD system. The OGOLOD dataset can be accessed through SPARQL or through a navigational interface (HTML/Linked Data).

- The first phase aims at allowing (semantic) web search engines to discover what is new or recently changed in our dataset in an efficient and timely manner. This phase includes two tasks: (1) to generate a set sitemap.xml files from our SPARQL endpoint, relying on automatic tools like sitemap4rdf [53]; (2) to submit the sitemap.xml files into (semantic) web search engines, such as Sindice [54] (We have already submitted our sitemap.xml [55] to Sindice).
- The second phase aims at including our dataset in the LOD cloud diagram (Fig. 9). To this end, we added an entry [56] of our dataset in the CKAN repository. The Linking Open Data Task Force provides some guidelines for collecting metadata on LD datasets in CKAN [57].

## 6. DISCUSSION

To the best of our knowledge, the OGOLOD system offers information currently absent in the LOD cloud: the relation between OMIM diseases and genes; the relation between OMIM diseases and human phenotypes from HPO; the relation between orthologs clusters and genes; the relation between genes, their GO annotations and the evidence codes of such relation. Therefore, the main contribution of the OGOLOD system is offering a one-stop service for accessing structured orthology/genetic diseases information and the context of such information from external resources.

Using the LOD approach to publish the OGOLOD dataset offers several advantages. By publishing structured information based on the RDF triple users can access the information using a well known query language like SPARQL, write applications that exploit our dataset and, even more, perform “massaging”, *i.e.* processing, on such information. Also, the connectivity of the data increases: the published information gets a dynamic context and other scientists can link to such information, favoring a light-weight and cost effective integration.

OGOLOD is designed to maximize the integration of its information with other datasets through:

- The use of RO, which works as a de facto standard for predicates in biomedical ontologies [58].
- The reuse of other OBO Foundry ontologies to enrich the OGOLOD ontology: Gene Ontology, Evidence Code Ontology, NCBI taxonomy and Human Phenotype Ontology.
- The integration of multiple repositories of clusters of orthologs: KOG, Homologene, OrthoMCL, and Inparanoid.
- The definition of links to other repositories of the life sciences LOD such as Bio2RDF and Linked Life Data.

Apart from the benefit of the integrated information, we have tried to follow LD best practices [33, 59] as much as possible in order to improve the user experience. For example, we have avoided the use of blank nodes in the case of evidence codes [60]. Vocabulary level links have also been added to the dataset to increase its integration potential [61].

The OGOLOD system is in its early development stages, and therefore there is room for improvement. In terms of the dataset itself, we are planning to use more vocabularies like Dublin Core [62] and other standard vocabularies, as well as improving the information about the graph [63]. We are also planning to improve the publishing workflow by incorporating tools like Pellet Integrity Constraint Validator [64] to check the consistency of the gathered information or SILK [65] to discover relationships between our dataset and external datasets. We are also planning to offer the possibility of using remote services for SPARQL querying [66].

In summary, albeit improvements are foreseen and necessary, OGOLOD is already a useful system in its current state since it offers an open platform of structured

orthologs/diseases information, built taking into account LD publishing best practices. OGOLOD lays the foundations for the development of other systems that can exploit our dataset via direct searches or software applications.

### ACKNOWLEDGEMENTS

This work has been funded by the Spanish Ministry for Science and Innovation through grant TIN2010-21388-C02-02. Mikel Egaña-Aranguren is funded by the Marie Curie-COFUND Programme (FP7) of the European Commission. Boris Villazón-Terrazas is funded by the Web n+1 (TSI-020301-2009-24) and BabelData (TIN2010-17550) projects of the Spanish Ministry for Science and Innovation.

### CONFLICT OF INTEREST

None declared.

### ABBREVIATIONS

LD	=	Linked Data
KB	=	Knowledge Base
LOD	=	Linked Open Data
OWL	=	Web Ontology Language
RDF	=	Resource Description Framework
SPARQL	=	Simple Protocol and RDF Query Language
URI	=	Uniform Resource Identifier
CKAN	=	Comprehensive Knowledge Archive Network
WWW	=	World Wide Web

### REFERENCES

- Shadbolt N, Berners-Lee T, Hall W. The SemanticWeb Revisited. *IEEE Int Syst* 2006; 21(3): 96-101.
- Semantic Web Health Care and Life Sciences (HCLS) Interest Group [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/2001/sw/hcls/>
- RDF Current Status [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/standards/techs/rdf>
- SPARQL Current Status [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/standards/techs-sparql>
- OWL Web Ontology Language Current Status [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/standards/techs/owl>
- URI Current Status [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/standards/techs/uri>
- Bizer C, Heath T, Berners-Lee T. Linked Data – The Story So Far. *Intl J Semantic Web Inform Syst* 2009; 5(3): -22.
- Linked Data [monograph on the internet]. Berners-Lee T; [cited 2011 Oct 22]. Available from: <http://www.w3.org/DesignIssues/LinkedData.html>
- Bizer C. Pay-as-you-go Data Integration on the public Web of Linked Data [monograph on the internet]. [cited 2011 Oct 22]. Available from: <http://www.wiwiss.fuberlin.de/en/institute/pwo/bizer/research/publications/Bizer-FIS2010-Pay-As-You-Go-Talk.pdf>
- Semantic Web stack image [homepage on the Internet]. Wikipedia; [cited 2011 Oct 22]. Available from: <http://en.wikipedia.org/wiki/File:Semantic-web-stack.png>
- Belleau F, Nolin M, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008; 41(5): 706-716.
- Antezana E, Kuiper M, Mironov V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief Bioinform* 2009; 10(4): 392-407.
- Smith B, Ashburner M, Rosse C, *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech* 2007; 25(11): 1251-1255.
- Miñarro Giménez JA, Madrid M, Fernández-Breis JT. OGO: an ontological approach for integrating knowledge about orthology. *BMC Bioinformatics* 2009; 10(10): S13+.
- Linking Open Data cloud diagram [homepage on the Internet]. Cyganiak R, Jentzsch A; [cited 2011 Oct 22]. Available from: <http://lod-cloud.net/>
- Tatusov R, Fedorova N, Jackson J, *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003; 4: 41-55.
- National Center for Biotechnology Information [homepage on the Internet]. NCBI; [cited 2011 Oct 22]. Available from: <http://www.ncbi.nlm.nih.gov/>
- Remm M, Storm CV, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001; 314(5): 1041-1052.
- Homologene [homepage on the Internet]. NCBI; [cited 2011 Oct 22]. Available from: <http://www.ncbi.nlm.nih.gov/homologene>
- Chen F, Mackey A, Stoeckert C, Roos D. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006; 34(1): D363-D368.
- Online Mendelian Inheritance in Man [homepage on the Internet]. OMIM; [cited 2011 Oct 22]. Available from: <http://www.ncbi.nlm.nih.gov/omim/>
- PubMed [homepage on the Internet]. NCBI; [cited 2011 Oct 22]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/>
- Egaña-Aranguren M, Stevens R, Antezana E, Fernández-Breis JT, Kuiper M, Mironov V. In: Knowledge-based Bioinformatics: From Analysis to Interpretation. Chapter: Technologies and best practices for building bio-ontologies. John Wiley & Sons Inc. 2010; pp. 68-86.
- Camon E, Magrane M, Barrell D, *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* 2004; 32: D262-D266.
- Evidence Codes Ontology [homepage on the Internet]. ECO; [cited 2011 Oct 22]. Available from: <http://www.obofoundry.org/cgi-bin/detail.cgi?id=evidencecode>
- Sayers EW, Barrett T, Benson DA, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2010; 38(1): D5-D16.
- Smith B, Ceusters W, Klagges B, *et al.* Relations in Biomedical Ontologies. *Genome Biology* 2005; 6: R46.
- Robinson PN, Khler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The Am J Human Genet* 2008; 83: 610-615.
- Keywords search interface [homepage on the Internet]. Miñarro-Giménez JA; [cited 2011 Oct 22]. Available from: <http://miuras.inf.um.es:9080/OGO-1.0/>
- Guided search interface [homepage on the Internet]. Miñarro-Giménez JA; [cited 2011 Oct 22]. Available from: <http://miuras-inf.um.es:9080/GuidedSearch/>
- Miñarro-Giménez JA, Egaña-Aranguren M, García-Sánchez F, Fernández-Breis JT. A Semantic Query Interface for the OGO Platform. Proceedings of the First international conference on Information technology in bio- and medical informatics; 2010 August 30 - September 3; Bilbao, Spain; pp. 128-142.
- Miñarro-Giménez JA, Miranda-Mena T, Martínez-Béjar R, Fernández-Breis JT. Exploitation of translational bioinformatics for decision-making on cancer treatments. Proceedings of the Second international conference on Information technology in bio- and medical informatics; 2011 August 29 - September 2; Toulouse, France; Springer 2011, pp. 1-15.
- Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. 1<sup>st</sup> ed. Morgan & Claypool 2011.
- Uniform Resource Identifier (URI): Generic Syntax [monograph on the internet]. Berners-Lee T, Fielding RT, Masinter L; [cited 2011 Oct 22]. Available from: <http://tools.ietf.org/html/rfc3986>
- The Data Hub - The easy way to get, use and share data [homepage on the Internet]. CKAN; [cited 2011 Oct 22]. Available from: <http://thedatahub.org>

- [36] Bio2RDF [homepage on the Internet]. Belleau F, Nolin M, Tourigny N, Rigault P, Morissette J; [cited 2011 Oct 22]. Available from: <http://www.bio2rdf.org/>
- [37] Linked Life Data [homepage on the Internet]. ONTOTEXT; [cited 2011 Oct 22]. Available from: <http://linkedlifedata.com/>
- [38] Punning [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/2007/OWL/wiki/Punning>
- [39] Orthology and Diseases Information: OGO [homepage on the Internet]. Miñarro-Giménez JA; [cited 2011 Oct 22]. Available from: <http://ckan.net/package/ogolod>
- [40] OGOLOD OpenLink Virtuoso SPARQL Query [homepage on the Internet]. Miñarro-Giménez JA; [cited 2011 Oct 22]. Available from: <http://miuras.inf.um.es/sparql>
- [41] Villazón-Terrazas B, Vilches L, Corcho O, Gómez-Pérez A. Methodological Guidelines for Publishing Government Linked Data. Springer 2011.
- [42] Cool URIs for the semantic web [monograph on the internet]. Sauer mann L, Cyganiak R; [cited 2011 Oct 22]. Available from: <http://www.w3.org/TR/2008/NOTE-cooluris-20080331/>
- [43] OpenLink Software [homepage on the Internet]. Virtuoso Universal Server. OpenLink; [cited 2011 Oct 22]. Available from: <http://virtuoso.openlinksw.com/>
- [44] Jena - A Semantic Web Framework for Java [homepage on the Internet]. HP; [cited 2011 Oct 22]. Available from: <http://jena.sourceforge.net/>
- [45] Sesame [homepage on the Internet]. Aduna; [cited 2011 Oct 22]. Available from: <http://www.openrdf.org/>
- [46] 4store [homepage on the Internet]. Harris S; [cited 2011 Oct 22]. Available from: <http://4store.org/>
- [47] YARS: Yet Another RDF Store [homepage on the Internet]. Harth A, Decker S; [cited 2011 Oct 22]. Available from: <http://sw.deri.org/2004/06/yars/>
- [48] OWLIM [homepage on the Internet]. ONTOTEXT; [cited 2011 Oct 22]. Available from: <http://www.ontotext.com/owlim>
- [49] Pubby - A Linked Data Frontend for SPARQL Endpoints [homepage on the Internet]. Cyganiak R, Bizer C; [cited 2011 Oct 22]. Available from: <http://www4.wiwiwss.fu-berlin.de/pubby/>
- [50] Joseki - A SPARQL Server for Jena [homepage on the Internet]. HP; [cited 2011 Oct 22]. Available from: <http://www.joseki.org/>
- [51] Talis Platform [homepage on the Internet]. Talis; [cited 2011 Oct 22]. Available from: <http://www.talis.com/platform/>
- [52] Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool 2011.
- [53] Sitemap4rdf. Cyganiak R, Villazón-Terrazas B; [cited 2011 Oct 22]. Available from: <http://lab.linkeddata.deri.ie/2010/sitemap4rdf/>
- [54] Semantic web index [homepage on the Internet]. Sindice; [cited 2011 Oct 22]. Available from: <http://sindice.com/main/submit>
- [55] OGOLOD sitemap [homepage on the Internet]. OGOLOD; [cited 2011 Oct 22]. Available from: [http://miuras.inf.um.es/sitemap\\_index.xml](http://miuras.inf.um.es/sitemap_index.xml)
- [56] Orthology and Diseases Information: OGO [homepage on the Internet]. CKAN; [cited 2011 Oct 22]. Available from: <http://ckan.net/package/ogolod>
- [57] Guidelines for Collecting Metadata on Linked Datasets in CKAN. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetadatainformation>
- [58] Beisswanger E, Lee V, Jung-Jae K, et al. Gene Regulation Ontology (GRO): design principles and use cases. *Studies Health Technol Inform* 2008; 136: 9-14.
- [59] Hogan A, Harth A, Passant A, Decker S, Polleres A. Weaving the Pedantic Web. Proceedings of the Linked Data on the Web Workshop (LDOW2010); 2010 April 27; Raleigh, North Carolina.
- [60] Blank nodes are forbidden [monograph on the internet]. Banff Manifesto; [cited 2011 Oct 22]. Available from: [http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff\\_Manifesto#Rule\\_234:\\_Blank\\_nodes\\_are\\_forbidden](http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto#Rule_234:_Blank_nodes_are_forbidden)
- [61] Vocabulary Links [monograph on the internet]. Heath T, Bizer C; [cited 2011 Oct 22]. Available from: <http://linkeddatabook.com/editions/1.0/#htoc21>
- [62] The Dublin Core Metadata Initiative [homepage on the Internet]. DCMI; [cited 2011 Oct 22]. Available from: <http://dublincore.org/>
- [63] Minimum information about a graph [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: [http://www.w3.org/wiki/HCLSIG\\_BioRDF\\_Subgroup/MinimalInformationAboutAGraph](http://www.w3.org/wiki/HCLSIG_BioRDF_Subgroup/MinimalInformationAboutAGraph)
- [64] Pellet Integrity Constraints: Validating RDF with OWL [homepage on the Internet]. Clark and Parsia; [cited 2011 Oct 22]. Available from: <http://clarkparsia.com/pellet/icv/>
- [65] Volz J, Bizer C, Gaedke M, Kobilarov G. Silk - a link discovery framework for the web of data. Proceedings of the 2nd Linked Data on the Web Workshop; 2009 April 20; Madrid, Spain; pp. 559-572.
- [66] SERVICE Graph Patterns [monograph on the internet]. W3C; [cited 2011 Oct 22]. Available from: <http://www.w3.org/TR/sparql11-federated-query/#service>