ORDO: An Ontology Connecting Rare Disease, Epidemiology and Genetic Data

Conference Paper · July 2014

CITATIONS
50

READS
1,759

9 authors, including:

Drashtti Vasant
Bayer HealthCare
9 PUBLICATIONS **957** CITATIONS
SEE PROFILE

James Malone
SciBite
84 PUBLICATIONS **5,493** CITATIONS
SEE PROFILE

Simon Jupp
European Molecular Biology Laboratory
74 PUBLICATIONS **2,738** CITATIONS
SEE PROFILE

Peter Robinson
Jackson Laboratory for Genomic Medicine, Farmington CT, United States
510 PUBLICATIONS **21,554** CITATIONS
SEE PROFILE

ORDO: An Ontology Connecting Rare Disease, Epidemiology and Genetic Data

Drashtti Vasant[1], Laetitia Chanas[2], James Malone[1], Marc Hanauer[2], Annie Olry[2], Simon Jupp[1], Peter N. Robinson[3], Helen Parkinson[1] and Ana Rath[2]*

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
[2]Orphanet – INSERM SC11, Plateforme Maladies Rares, 96, rue Didot, Paris 75014, France
[3]Institute for Medical Genetics, Charitè-Univerisitätsmedizin Berlin, 13353 Berlin, Germany

ABSTRACT

Motivation: Orphanet serves as a reference portal for rare diseases populated by literature curation and validated by international experts. The Orphanet information system is supported by a relational database designed around the concept of a disorder. Increasingly, Orphanet is seen as a reference for this domain and as such is required for reuse by external applications. These applications require complex queries, specific views tailored to user groups or investigation areas and integration or cross-referencing with resources such as OMIM and Ensembl. A formal, portable and open-access ontological representation of Orphanet is required by the community.

Results: We present the Orphanet Rare Disease Ontology (ORDO), an open-access ontology developed from the Orphanet information system, enabling complex queries of rare disorder and its epidemiological data (age of onset, prevalence, mode of inheritance) and gene-disorder functional relationships. Bespoke views can be extracted using the ontology axiomatisation eg. phenome-disorder views.

Availability: ORDO (OWL and OBO format) is available http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php. ORDO can be browsed in BioPortal (http://bioportal.bioontology.org/ontologies/ORDO) and in OLS (Ontology Lookup Service) (https://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=Orphanet).

1 INTRODUCTION

Historically, there has been a shortfall of medical and scientific knowledge in the field of rare diseases primarily due to a lack of funded research or public health policy (Tambuyzer, 2010). In Europe, a disease that affects 1 in every 2,000 people is considered rare (Rath et al., 2012) and the ODA (Orphan Drug Act) defines rare diseases as those affecting fewer than 200,000 people in the United States (Tambuyzer, 2010). Orphanet has maintained a reference portal since 1997 and provides access to information about rare diseases and orphan drugs - those specifically developed to treat rare disorders. Orphanet is a "disorder" centric resource, in contrast to Online Mendelian Inheritance in Man (OMIM) (McKusick, 1998), which defines entries on their genetic basis. The portal is supported by a multilingual database which is populated by literature curation and validated by international experts (Rath et al., 2012). To date more than 7,000 disorders are included in the Orphanet database and new disorders are added regularly. The database integrates (in a number of languages) the nosology (or classification) of rare diseases, their relationship with genes and epidemiological data, cross-references to other terminologies, databases and classifications.

Increasingly, Orphanet is seen as a reference for this domain and as such is required for reuse by external applications. These applications include complex queries such as all disorders of a phenome and with a specific mode of inheritance and defined age of onset. In addition views tailored to user groups or investigation areas are required. This requires the ability to filter or include particular biological entities depending upon a given criteria and to manage the poly-hierarchies which are introduced by addition of these views, for example, provide 'all the disorders which are morphological anomalies'. Finally, interoperability with resources such as OMIM and Ensembl is important to provide links to genetic disease.

We have developed the Orphanet Rare Disease Ontology (ORDO). ORDO is a portable and open-access representation of the data in the Orphanet information system, formalised as an OWL ontology. The ontology includes Orphanet concepts as OWL classes including phenomes, disease, genes, genetic inheritance mode and prevalence. We use OWL to explicitly model relationships between these classes and, by the use of inference through description logic reasoning, enable powerful querying This satisfies two of our requirements, provision of complex queries across the resource and the ability to generate specific views or create and manage a poly-hierarchy (Jupp et al., 2012). An additional benefit is that any logical inconsistencies in the data are detected by automated inferencing over ORDO aiding in

* To whom correspondence should be addressed.

knowledge management (addition, curation, validation and quality control) in each release (Rath et al., 2012).

ORDO has been applied to database resources including ArrayExpress, BioSamples, Ensembl and the Gene Expression Atlas all of which use the Experimental Factor Ontology (Malone et al., 2010) and which imports the rare disease classification hierarchy from OMIM. ORDO is available from BioPortal, Ontology Lookup Service and directly from the Orphanet website.

2 METHOD

Orphanet provides a database export as XML files containing a subset of the data (http://www.orphadata.org/cgi-bin/index.php). These datasets (see Table 1) were used in the construction of ORDO. A freely available ontology generation tool (https://github.com/Orphanet/Orpha2Ordo/tree/master/OrphoToOWL) downloads the latest XML (Table 1) from the Orphanet website, a series of ontology design patterns are applied (Figure 1) and the XML is translated into an OWL and OBO file. The process is run monthly and new releases of ORDO are generated in both OWL and OBO formats.

Table 1: Orphanet XML files used as the basis of ORDO. Each file defines a unique set of entities.

File	Entity	Example
http://www.orphadata.org/data/xml/en_product1.xml	Rare Disorder Label, Synonym, Cross-references, Phenome Type	'Hereditary angioedema type 1 ' 'HAE-1' 'OMIM:106100' 'etiological subtype'
http://www.orphadata.org/data/xml/en_product2.xml	Age of Onset, Mode of Inheritance, Prevalence	'unknown' 'autosomal dominant' '1-9 /100,000'
http://www.orphadata.org/cgi-bin/inc/product3.inc.php	Classification of Rare Diseases,	'child of hereditary angioedema'
http://www.orphadata.org/data/xml/en_product6.xml	Gene Label, Gene-disorder Relation, Gene Xref, Gene Synonyms,	SERPING1 'disease-causing germline mutation in' 'HGNC:1228' 'plasma protease C1 inhibitor'

Orphanet contains a hierarchical clinical classification of rare disorders; which is organized into medical specialities such as rare genetic disorders, rare cardiac disorders etc) (Rath et al., 2012). Orphanet also assigns each disorder a phenome type. Phenome is defined as 'a set of phenotypes expressed at the cell, tissue, organ or organism level. It describes the "physical totality of all traits of an organism or of one of its subsystems"'. Phenome types are listed in Table 2 each of these is a unique OWL class in ORDO.

Table 2: Orphanet phenome-types assigned to rare disorders

Phenome type	Example
biological anomaly	Methylmalonic aciduria due to transcobalamin receptor defect
clinical subtype	Adult Krabbe disease
clinical syndrome	Meigs syndrome
etiological subtype	African tick typhus
group of disorders	Rare bone disease
histopathological subtype	Ependymoma
disease	Acatalasemia
malformation syndrome	Ackerman syndrome
morphological anomaly	Anodontia

An example of this two-tier classification of rare disorders is: retinoblastoma is-a rare eye tumor (clinical specialty) and Retinoblastoma is-a disease (assigning the phenome-type). ORDO models both of these assigning explicit relationships (is-a and part-of) between the disorders.

Table 3: Example complex queries of ORDO in natural language and Manchester syntax.

Query	Manchester OWL Syntax
a) Query for all rare genetic bone diseases that have the age of onset as Neonatal/infancy and range of prevalence is 1-9/1,000,000.	*Rare genetic bone disease' or (part_of some 'Rare genetic bone disease') and has_prevalence some '1-9 / 1,000,000' and has_AgeOfOnset some Neonatal/infancy.*
b) Query for all genes with disease-causing germline mutations in some morphological anomaly where morphological anomaly has mode of inheritance autosomal recessive.	*gene and 'Disease-causing germline mutation(s) in' some ('morphological anomaly' and (has_inheritance some 'autosomal recessive')).*

The ontology was produced with a set of competency questions, (Table 3) used to guide the development and assess

the resulting ontology. To fulfill these queries, explicit relationships were defined (see section 3) with between various ontology classes.

3 RESULTS

ORDO consists of 11,699 classes and 76,554 annotation assertion axioms represented in OWL using the modeling schema shown Figure 1.

Each concept from the Orphanet database forms a distinct OWL class and is associated with other classes using a set of defined object properties. Since all the phenome subclasses are disjoint (i.e. a disorder cannot be a clinical subtype and a clinical syndrome at the same time and so on), a part_of relationship was used to assert the classification when:

if (Disorder A phenome_type) != Parent (Disorder A) phenome_type)

For example, *Familial lambdoid synostosis (morphological anomaly) is_a Isolated craniosynostosis(group of disorder)* and *Familial lamboid synostosis* ***part_of*** *Isolated craniosynostosis.*

ORDO also represents the relationship between the disorders and their genetic cause (if known), the mode of inheritance and associated epidemiological data (age of onset,

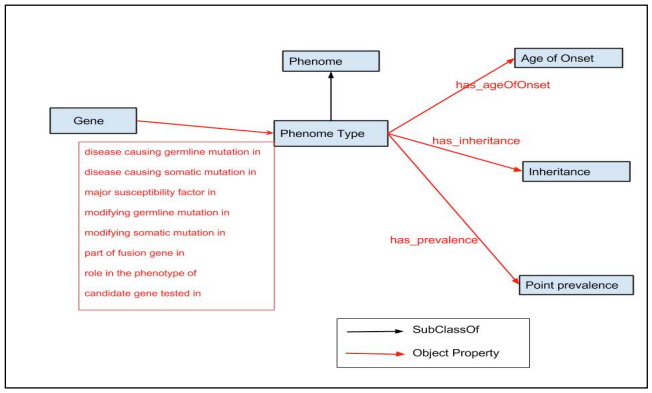


Figure 1: Modeling schema adopted by ORDO

age of death, prevalence) as seen in Figure 1, and not just the nosology such as that captured in the Disease Ontology (Schriml et al., 2012).For eg. Class: *Tibia hemimelia* is described as a *SubClassOf morphological anomaly, has_inheritance some sporadic, has_AgeOfOnset some neonatal/infancy, has_prevalence some 1-9/1000,000 and part_of some Hemimelia.* This is an important distinction and this information is of value in the drug discovery process and when performing genetic diagnostics of undiagnosed disorders e.g. by exome sequencing.

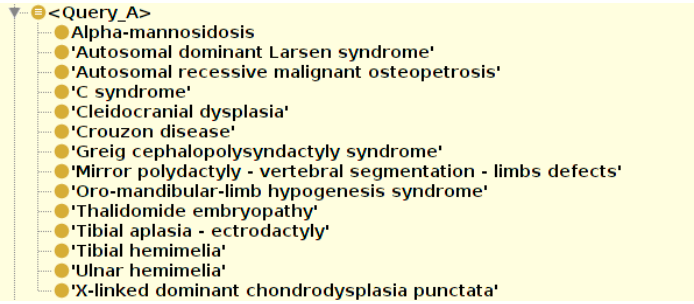
Each class is also associated with annotations such as label, alternative term and cross-references. The Evidence Code Ontology (ECO) (Karp et al., 2004), is also used to encode the provenance of assertions made in ORDO. For example, the gene *ADAMTS-like 4* is asserted to contain 'Disease-causing germline mutation(s) in' some 'Isolated ectopia lentis'; this assertion is annotated with ECO:0000205 or curator inference with the value “Curated” indicating this assertion was curated or confirmed by an expert curator.

ORDO also provides disease cross references to the International Classification of Diseases (10th version), SNOMED-CT, MeSH, MedDRA, OMIM and UMLS (Bodenreider, 2004) and genes are cross-referenced to HGNC (Povey et al., 2001), UniProt (UniProt Consortium, 2008), OMIM, Ensembl (Flicek et al., 2014), Reactome (Matthews et al., 2008) and Genatlas (Frezal, 1998). For example, hereditary angioedema is mapped to OMIM:106100 (angioedema, hereditary,type 1;HAE1) and ICD10:D84.1(Angioedema, hereditary). These mappings are reviewed for accuracy by experts and this enables wider data integration with other resources increasing domain interoperability and providing a classification of rare disease accessible to resources e.g. those cross referenced to OMIM. It is important to note that the cross-references between ORDO and OMIM are not one-to-one, as the granularity and organisation of the respective resources are different.

3.1 Querying ORDO

The use of class descriptions in OWL described here, enable more complex querying which was difficult or impossible using the existing relational database. Using the same examples as the methods section, queries were run using the defined classes shown in Table 3. The results of these are shown in Figures 2 and 3 respectively. The class Tibia hemimelia described before will now appear while running Query_A in Figure 2. Although these defined classes are not included within ORDO the use of OWL axiomatisation makes the addition structure possible and allows users to add these as needed. ORDO therefore provides means of automated inference, validation and curation of data and provides a new and richer mode of access than previously possible.

Figure 2: Query result -14 classes- for all rare genetic bone diseases that have the age of onset as Neonatal/infancy and its range of prevalence is 1-9/1,000,000 as visualised in Protégé 4.



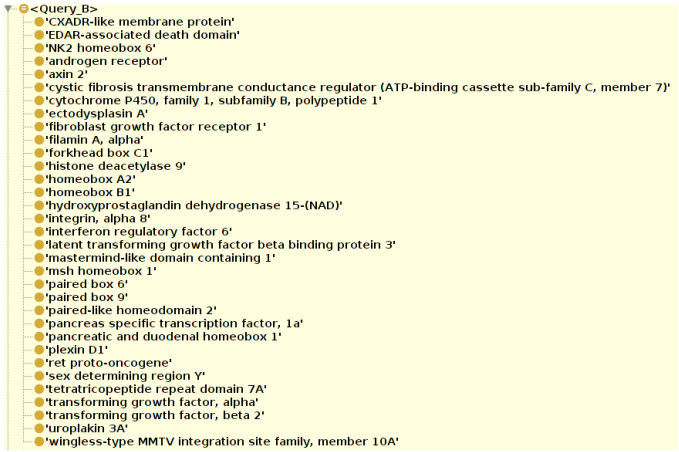


Figure 3: Query result - 33 classes - for all genes that are disease-causing germline mutations in some morphological anomaly and that morphological anomaly has mode of inheritance autosomal recessive as visualised in Protégé 4.

4 DISCUSSION

The organisation of disorders based on their phenome type, is of value to the scientific community as it offers the possibility of improving phenotypic models for further research (Pouladi et al, 2013). A module containing the "Rare genetic disorder" branch of ORDO is automatically extracted for each release and imported into the Experimental Factor Ontology (EFO), an data driven application ontology. EFO is used by several resources (ArrayExpress, Ensembl, PRIDE etc) within EBI and external projects. By inclusion of this ORDO import our resources can be queried for all rare disorders. EBI is now exploring disease and phenotype content across its resources and in future by use of ORDO will organise both common and rare diseases improving query results and the search experience for users. We will also use ORDO in the International Mouse Phenotyping Consortium portal (www.mousephenotype.org) to integrate mouse models of disease annotated with OMIM identifiers and in the annotation of Induced Pluripotent Stem Cell lines derived from rare genetic disease patients by the HIPSCI project (www.hipsci.org/)to integrate molecular data deposited by the project in EBI's databases. In the future, we will to enrich the ontology in future by inclusion of more information about each disorder. For example, average age of death for the disorder, prevalence and incidence figures by country/population and whether the disorder is caused by a loss or gain of gene function. Efforts are also underway to integrate ORDO with the Human Phenotype Ontology (Robinson & Mundlos, 2010) annotating Orphanet's phenome types with appropriate HPO terms. This will provide interoperability between projects such as RD-Connect and Decipher which use the HPO and will drive the revision of the phenome hierarchy once HPO terms have been integrated. Requests for ontology edits and new terms can be made via https://www.ebi.ac.uk/panda/jira/browse/ORDO/. Users should subscribe to ORDO announce to be informed of new releases https://listes.inserm.fr/sympa/info/ordo-users.orphanet.

ACKNOWLEDGEMENTS

This work is funded in part by EMBL-EBI core funds, MRC/Wellcome Trust Strategic Award HIPSCI and Inserm, French Directorate General for Health, the European Commission and the and the Bundesministerium für Bildung und Forschung (BMBF project number 0313911).

REFERENCES

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267–70.

Flicek, P. et al. (2014) Ensembl 2014. *Nucleic Acids Research*. 42 Database issue:D749-D755

Frézal, J. (1998). Genatlas database , genes and development defects. *Elsevier*, *321*(10), 805–817.

Jupp, S., Gibson, A., Malone, J., & Stevens, R. (2012). Taking a view on bio-ontologies. In ICBO 2012, Graz.

Karp, P. D., Paley, S., Krieger, C. J., & Zhang, P. (2004). An evidence ontology for use in pathway/genome databases. *Pacific Symposium on Biocomputing,*190–201.

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., … Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics (Oxford, England)*, *26*(8), 1112–8.

Matthews L. et al. (2008) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*. 2008;37 Suppl 1:D619-D622.

McKusick, V.A.: Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 1998 (12th edition)

Pouladi, M. a, Morton, a J., & Hayden, M. R. (2013). Choosing an animal model for the study of Huntington's disease. *Nature Reviews. Neuroscience*, *14*(10), 708–21. doi:10.1038/nrn3570

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., & Wain, H. (2001). The HUGO Gene Nomenclature Committee (HGNC). *Human Genetics*, *109*(6), 678–80.

Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., & Ayme, S. (2012). Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Human Mutation*, *33*(5), 803–8.

Robinson, P. N., & Mundlos, S. (2010). The human phenotype ontology. *Clinical Genetics*, *77*(6), 525–34.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., … Kibbe, W. A. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, *40*(Database issue), D940–6. doi:10.1093/nar/gkr972

UniProt Consortium. (2008). The universal protein resource (UniProt). *Nucleic Acids Research*, *36*(Database issue), D190–5.

Tambuyzer, E. (2010). Rare diseases, orphan drugs and their regulation: questions and misconceptions. *Nature Reviews. Drug Discovery*, *9*(12), 921–9.