Registries of domain-relevant semantic reference models help bootstrap interoperability in domains with fragmented data resources

Marco Roos[1], Mark D Wilkinson[2], Rajaram Kaliyaperumal[1], Mark Thompson[1], Claudio Carta[3], Ronald Cornet[4,5], David van Enckevort[6], and Luiz Bonino[7]

[1] Leiden University Medical Centre, The Netherlands
{m.roos,r.kaliyaperumal,m.thompson}@lumc.nl
[2] Universidad Politcnica de Madrid, Spain
markw@illuminae.com
[3] Istituto Superiore di Sanitá, Italy
claudio.carta@iss.it
[4] Academic Medical Center University of Amsterdam, The Netherlands
[5] Linköping University, Sweden
r.cornet@amc.uva.nl
[6] University Medical Center Groningen, The Netherlands
david.van.enckevort@umcg.nl
[7] Dutch Techcentre for Life Sciences, The Netherlands
luiz.bonino@dtls.nl

Abstract. The specialist field of rare diseases must connect its vast array of globally distributed disease and patient registries to maximise their value. Unfortunately, many registries are "boutique", with few or no staff with formal informatics training. At a series of Bring Your Own Data workshops, we helped registry owners transform their data into formally structured triple stores following the Linked Data principles and demonstrated the potential of data linkage. We documented several useful approaches that we believe could be followed independently by other registry owners worldwide, including: that the transformation to Linked Data could be considered as passing through layers of increasing semantic complexity; that only a subset of ontologies are relevant at each layer; and that certain data transformation processes could be modelled as an "archetype", and presented to registry staff to fill-in with their data. We propose that formally capturing these ontological layers and archetypes, and registering them as a reference and teaching resource will facilitate the wider community of non-expert data owners self-direct their own data transformations.

Keywords: semantic reference model, linked data, ontologies, rare disease, archetype

1 Introduction

Making data linkable at the source has become a key ambition for the development of robust infrastructure that supports data integration in the rare disease community. There are over 6000 rare diseases, each with multiple data resources across the globe, ranging from biobanks, patient or disease registries, and omics data sources. We must accept the challenge of implementing solutions that can scale-up to be adopted by thousands of resources, with the knowledge that maintaining a centralised warehouse at this scale, and with this kind of sensitive data, is neither feasible nor ethically or legally acceptable.

Perhaps more than in other domains, progress in the rare disease field depends on combining data, given that the disease-specific data is so sparse. It is, however, well established that biomedical data integration is an extremely error-prone process that requires a deep understanding of both biology and data/knowledge management to reconcile data from different sources. To improve the data ecosystem for this important target community, we are working on a standard set of procedures and lightweight technologies that will make rare disease data Findable, Accessible, Interoperable and Reusable for both humans and computers (FAIR) at the source.

In this position paper we discuss requirements and subsequent design decisions that we have chosen to pursue during a still ongoing plan to make rare disease biobanks and registries linkable at the source. The plan also includes a study of the steps to make data FAIR at the molecular level (e.g. genetic variants, metabolites, molecular pathways) in order to link to information in registries and biobanks. The plan is guided by experiences gained from a number of Bring Your Own Data workshops (BYODs) in the rare disease domain [14, 13]. While not all components described here have been built or tested, we take the position that our early successes in the early stages of this approach suggest that the future extensions - currently in development and based on the same layered design - will exhibit similar successes.

2 Backbone: Linkable Data and Ontologies

Choosing Linked Data principles and Ontologies to make rare disease data linkable at the source was our first design decision, as RDF was designed with the objective of creating qualified networks of data, upon which increasingly complex domain models can be overlaid to assist with interpretation of that data. For instance, the Human Phenotype Ontology and the Orphanet Rare Disease Ontology are obvious choices to denote human phenotypes and diseases in rare disease resources within this Linked Data. We therefore considered this the best way to facilitate integrative biological and translational research across rare disease resources. Other tools in this general domain that use ontologies include the exomizer, matchmaker exchange tools, and Monarch, providing examples of the power of using phenotype annotations and cross-species phenotype mappings [6]. RDF is capable of representing disease specimen identifiers, patient/disease

personal and clinical information, and molecular data, thus the choice of this singular technological framework helps reduce the overall cost of downstream data integration for rare disease resources. As such, we received wide support for putting this first design decision into practice from many sources, such as RD-Connect, Elixir, BBMRI, ODEX4All, FAIRDict, academic hospitals, and an increasing number of patient organisations.

2.1 Composite semantic models as reference for preparing data for integration

Our position is that, by making an explicit set of increasingly rich semantic layers (diagrammed as a set of "Modules" in Figure 1), where each Module may be taught and undertaken in-isolation from the others, focuses trainees on the specific subset of tasks and ontologies required to achieve success in that layer. We will now elaborate on that position.

Linked Data with strong ontological underpinnings, and a clear model for achieving proper access control, was our first ambition for preparing the relatively small, but numerous and disparate, rare disease data sets for wide-scale data integration. However, an immediate and major bottleneck was the sparsity of expertise in the community to make informed decisions about which ontological concepts to use for their data annotations. Searching for a concept, e.g. in NCBOs bioportal or EBIs ontology lookup service, typically returns too many hits for a non-ontologists to choose from. Specific ontologies may be advised by experts, but the breadth of data types across data sets is large. For example, working with rare disease patient registry managers, we easily listed at least 10 ontologies relevant for even a small a subset of their registry's data, and not all of these are included in the BioPortal or EBI search services. Providing our target community with too many choices will be confusing. At the same time, investigating individual ontologies for each rare disease resource that we prepare for analysis across data sets is time consuming and inefficient.

Ideally, therefore, we should attempt to record and reuse previous ontology-assessments every time we go through the process of making a rare disease resource linkable, such that we consistently advise only one or at most a small number of ontologies for any given class or type of data/observation. The key objective of Module 1, therefore, was to create a searchable subset of domain-relevant ontological resources or ontology-slices, rather than asking our community to do an open ontology search for every term. This provides a way for non-experts to start to become good Linked Data publishers and reduces confusion and frustration for our rare disease registry community. Moreover, because the constraints are only on what we present to the data publisher, the power of the full ontology remains available to machines that consume or query that data. To date, we have successfully used these approaches to assist a number of rare disease data registries - many of them with little or no formal training in data or knowledge management, to create Linked Data from their data that is of sufficiently high quality that it can be used as a source for federated SPARQL

queries. We now wish to scale-up these efforts, such that registry owners require ever-fewer formal contacts with Linked Data or ontological experts.

The next layer in our stack, Module 2, guided again by our experience working with this community, derives from our observation that many of the core data models from patient registries and biobanks have near-identical structures, particularly for similar 'types' of data (for example, clinical observations are similar in structure to each other, but distinct from coded phenotypic observations). We have therefore started to compose semantic reference models for rare disease data integration (closely related to Archetypes in health information systems, e.g. see [5]) that our community can simply copy and populate with their specific data. These, too, will be published in a searchable registry of such models, and will be 'tagged' with keywords related to the kinds of data they are capable of representing. Note that these models are data *structures*, not novel semantic models of diseases or phenotypes - we do not intend these models to be new conceptualisations of a domain - in the sense that, for instance, the Human Phenotype Ontology is a distinct conceptualisation in the phenotypic domain; rather, these are meant as artefacts to further our data integration goals, which, together with the constrained ontological choices, suggest/limit both structure and semantics, reducing freedom-of-choice, but enhancing interoperability through capturing what we believe are the best-practises defined by data publishing experts. Archetypes are initially being designed through a collaboration between rare disease domain experts and Linked Data experts until a mutually-acceptable model is created. This model will then be published as a reference for individual data owners to build Linked Data within their domain/scope. Through our ongoing pursuit this approach, we intend to gradually build-up a clearly-defined set of starting points for all of the various data-types in the rare disease domain, allowing us to rapidly scale-up to absorb new resources into our integrated community through their own individual efforts.

This stack is currently being extended further (Module 3), where we are planning to create resources of archetypes with greater semantic complexity that may be used to combine individual local observations, as was done in [15], or link local observations with remote observations. As we create these reference models, we propose to distinguish between the three distinct outcomes that the models should support. These are, in order of complexity: the need to (i) harmonize and simplify annotation of source data, (ii) query across resources and enable statistical analysis of knowledge graphs, and (iii) enable logical reasoning to facilitate discovery by revealing"unknown unknowns". We propose to follow the modelling suggestions of the SemanticScience Integrated Ontology [4] that semantic models defined using OWL axioms allow a modular, layered approach, resulting in a composite model that can address each of these needs. Thus each of our proposed modules is an independent OWL file that can be utilised in isolation, depending on the expertise of the publisher.

Each module in the stack serves a specific, and increasingly more complex integrative purpose; the full spectrum of requirements - up to and including semantic reasoning - will only be achieved when all of the modules have been used

to annotate/represent the data. For instance, Module 1 (green, in Figure 1) for core identifier annotation primarily recommends ontological classes that allow explicit typing of the identifiers commonly seen in rare disease databases, but excludes any deeper properties such as those that would facilitate faceted data integration or querying based on properties or their values. In our current model we use the EMBRACE Data and Methods model (EDAM [7]) as the source of identifier-type semantics, and we encourage the use of the identifiers.org URI schemes to harmonise the identifier structures, where appropriate. The application of Module 1, therefore, facilitates simple queries to identify repositories that contain data of a particular nature, but are insufficient for the more complex integrative behaviours. Such richer behaviours are enabled by applying the archetypes and ontologies recommended by Modules 2 and 3. What is important to note is that the layers separate and stratify the tasks of semantic data migration. Module 1 starts with the most core question "what data do I have", which is in-itself an important piece of semantic information. The layers make it clear that this basic task can and should be clearly separated from the other, more complex tasks required to support full integrative queries.

We argue that the task undertaken in Module 1 is sufficiently comprehensible and self-evident that it provides non-ontologists an relatively easy way to pursue semantic transformations unaided by a data linking expert. We take the position that the "shallow" semantic transformation undertaken by Module 1 not only provide useful integrative behaviours, but do not in any way compromise the later addition of greater semantic expressivity, guided by Modules 2 and 3. We further hold the position that both Module 1 and Module 2, when presented to the community as a limited set of choices, provide a level of expectation well-within the capabilities of our target, non-expert data publising community. While the complexity of ontologies is a bottleneck for many who approach semantic transformations on their own, we propose that this layered approach lowers the bar for participation, and will stimulate more registries to undertake these preliminary transformations "at-source", on their own initiative.

As mentioned, Modules 2 and 3 are still under-development. We believe that these Modules will contain recommendations that can be reused to stimulate interoperability between resources. They provide predicates that define the relations between individuals and their observed or measured clinical features/phenotypes, and archetypes for how to assemble these observations without loss of data. For example, using the Semanticscience Integrated Ontology model (SIO; [4]) for recording measurements ensures that all clinical observations must include a value, a measurement unit, and an ontological type (for example, systolic blood pressure). Tables and relational database models rarely explicitly express such semantics, and thus by providing these simple, but rigorous archetypes, we provide a clear path forward for those who wish to further transform their data. Moreover, by agreeing on archetypes, we are able to create registry software that uses these as data-capture templates, ensuring that newly generated data fills these richer models without requiring extensive training of the registry owners. Indeed, we are working with patient registry software

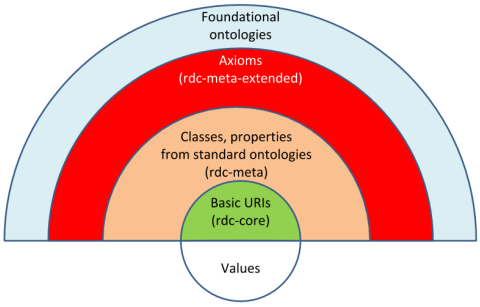

Fig. 1. Stacked modules (owl files) of the semantic reference model for increasingly complex cases. 'Values' represent data in multiple resources ; module 1 (green): simple classes for database identifiers; module 2 (light-red): immediately relevant classes and properties to denote the meaning of identifiers and their interlinks ; module 3 (red): axioms from the reused ontologies needed for reasoning; top module (light-blue): foundational ontologies that the reused ontologies refer to.

providers to incorporate support for this directly in their tools. For existing data, we can apply tools such as OpenRefine with the RDF plugin to add URIs for data values (e.g. HPO URIs for phenotypes) and data types, their interrelations, and their links to the reference model.

2.2 A prototype semantic reference model for enabling questions across rare disease resources

We have created a first version of a semantic reference model in the rare disease domain. Its main purpose is to enable questions across rare disease biobanks and registries. This is reflected by separate modules that comprise our reference model (figure 1), an example is given in figure 2. Each module is available as a separate owl file ([8]). The model refers to EDAM [7] for identifiers, OBIB (Ontology for Biobanking [3]) for biological specimen, ORE (the Object Reuse and Exchange model [9]) for aggregating research materials (a decision inspired by the research object model [1, 2]), and assumed the use of HPO (the Human Phenotype Ontology [12]) for phenotype identifiers and ORDO (Orphanet Rare Disease Ontology [17, 8]) for rare disease identifiers. At this time, we make no further assumptions as to which ontologies to recommend in the rare disease domain, but this is anticipated with support from RD-Connect. We included some initial mappings to for instance SNOMEDCT [16]. With our collaborators, we have also started work on mappings to MIABIS (Minimum Information About

[8] https://github.com/LUMC-BioSemantics/Rare-Disease-Semantic-Model

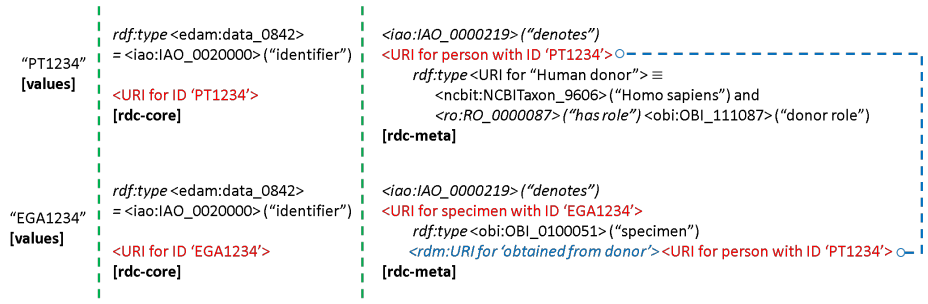

Fig. 2. Example data and pseudo-rdf from different modules (bold). The broken green lines indicate the crossing between modules. The blue dotted line indicates the semantic link between data elements. Prefixes indicate the reuse of entities from standard ontologies, for which the axioms can be found in rdc-meta-extended and upper ontologies (not shown). No prefix indicates that a URI was minted within the rdc namespace.

BIobank data Sharing [11]). In a next revision, we also aim to address the issue of implicit reinterpretation of rare disease data by changes in the ontologies that we use in the reference model. Assuming that ontology versioning is still imperfect, this may require an additional layer in our stacked approach. The current model is stored in github, and we have created an entry in BioSharing[9]. We have not (yet) decided on uploading the model to NCBOs BioPortal.

3 Registries for domain-specific semantic reference models

Composite semantic reference models provide a useful service for data integration. In our experience however, these models could be better supported. Considering the FAIR paradigm, they are currently hard to make findable, accessible and reusable. Model registries and lookup services could make them more findable as data integration artefacts, and advocate them as standard schemas for data annotation in specific domains (in our case the rare disease domain). We envision searching for models for data integration by linked data graphs, such as a search for models that prepare data for linking genes to diseases (*'Which model can make genes in my data set linkable to diseases in other people's data?'*). We have uploaded our alpha version to BioSharing[9], which may be the appropriate platform. It aims to be a central point to find standards, and it will help reuse because we can add example annotated data. We envision that disease and sample registry software will use semantic archetype registries to optimize data entry towards generating interoperable data. However, at this time it has no special features for searching semantic models that prepare a resource for data integration. Tools such as EBIs Zooma and Ontology Lookup Service help users find a wealth of possibilities with great precision, but do not yet allow filtering

[9] https://biosharing.org/bsg-s000676

on semantic archetypes to limit the search results. For example Zooma returns a list of concepts for the search terms and shows which other sources use that particular ontological term, but not the archetype of the source itself.

4 Discussion

The concept of building composite models from existing ontologies for specific applications is not new, and they often help data integration. For example, EBIs Experimental Factor Ontology was developed as an application ontology for linking EBI [10] resources, and the Just Enough Results Model [18] may be considered a reference model for linking systems biology data within the FAIRdom initiative[10]. The need to simplify the ontological landscape for applied ontologists also seems a strong motivation for developing SIO [4]. Our suggestion is to mitigate this need by reusing the work of ontologists through published data integration models that apply state of the art axiomatized ontologies.

We take the position that it is useful to create a registry that supports finding, accessing and reusing domain-relevant subsets of ontological classes and Linked Data models. When designing this approach, we took the position that the task can be cast into three independent Modules that address distinct levels of semantic complexity; we hope that semantic model tool builders will now investigate how well their tools support this. Finally, we take the position that, when cast in this way, the tasks represented by Modules 1 and 2 become tractable to non-experts in data publishing, due to the enhanced clarity and simplified, task-specific, search results.

5 Conclusion

We are creating a reusable semantic reference model to speed up the process of making rare disease data resources FAIR and linkable at the source. This pertains to the many patient/disease registries, biospecimen collections (biobanks), and omics data resources that we need to be able to query across in order to speed up rare disease research in healthcare and life science. We have observed in previous BYODs that finding recommendable concepts in existing ontologies is the main bottleneck for rare disease stakeholders, and a redundant time investment for Linked Data experts. We take the position that we can mitigate this by improving support specifically for semantic models that are made to facilitate data integration downstream of semantic data encoding and annotation. These semantic models should be easy to find, access, and reuse. For interoperability use cases beyond findability, we advocate a modular approach, providing appropriate modules for data annotation, enabling simple manual queries, and big data analytics and reasoning. We propose that BioSharing could be a target for extending support for semantic data integration models, for instance by allowing searches for linked data patterns.

[10] http://fairdom.org

Acknowledgments. We thank all domain experts and linked data experts who contributed to previous Bring Your Own Data workshops for rare disease registries and biobanks. We thank the colleagues at ISS (Istituto Superiore di Sanità, Rome, Italy), particularly Sabina Gainotti and Domenica Taruscio, and Mascha Jansen (Dutch Techcentre for Life Sciences) for their preparatory work in organising these workshops. We thank Andrew Gibson and Katy Wolstencroft for fruitful discussions. The work leading to this paper is supported by grants from RD-Connect (FP7/20072013, grant agreement No. 305,444), Elixir infrastructure for life science data, Elixir-Excelerate (H2020-INFRADEV-1-2015-1), and BBMRI-NL2 (NWO National Roadmap for Large-Scale Research Facilities). MDW is supported by the Fundacion BBVA and the UPM Isaac Peral programme, and the Spanish Ministerio de Economa y Competitividad grant number TIN2014-55993-R.

References

1. Bechhofer, S., Roure, D.D., Gamble, M., Goble, C., Buchan, I.: Research Objects: Towards Exchange and Reuse of Digital Knowledge (feb 2010), http://eprints.ecs.soton.ac.uk/18555/
2. Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J.M., Bechhofer, S., Klyne, G., Goble, C.: Using a suite of ontologies for preserving workflow-centric research objects. Web Semantics: Science, Services and Agents on the World Wide Web 32, 16–42 (may 2015), http://linkinghub.elsevier.com/retrieve/pii/S1570826815000049
3. Brochhausen, M., Zheng, J., Birtwell, D., Williams, H., Masci, A.M., Ellis, H.J., Stoeckert, C.J., Jr.: OBIB-a novel ontology for biobanking. Journal of biomedical semantics 7, 23 (2016), http://www.ncbi.nlm.nih.gov/pubmed/27148435http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4855778
4. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., Klassen, D., McCusker, J.P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M.D., Hoehndorf: The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. Journal of Biomedical Semantics 5(1), 14 (2014), http://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-14
5. Ellouze, A.S., Bouaziz, R., Ghorbel, H.: Integrating semantic dimension into openEHR archetypes for the management of cerebral palsy electronic medical records. Journal of Biomedical Informatics 63, 307–324 (2016)
6. Haendel, M.A., Vasilevsky, N., Brush, M., Hochheiser, H.S., Jacobsen, J., Oellrich, A., Mungall, C.J., Washington, N., Köhler, S., Lewis, S.E., Robinson, P.N., Smedley, D.: Disease insights through cross-species phenotype comparisons. Mammalian Genome 26(9-10), 548–555 (oct 2015), http://link.springer.com/10.1007/s00335-015-9577-8
7. Ison, J., Kalas, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., Rice, P.: EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics (Oxford, England) 29(10), 1325–32 (may 2013), http://www.ncbi.nlm.nih.gov/pubmed/23479348http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3654706

8. Kahn, C.E.: Integrating ontologies of rare diseases and radiological diagnosis. Journal of the American Medical Informatics Association 22(6) (2015)
9. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S., Sanderson, R., Johnston, P.: A Web-based resource model for scholarship 2.0: object reuse & exchange. Concurrency and Computation: Practice and Experience 24(18), 2221–2240 (dec 2012), http://doi.wiley.com/10.1002/cpe.1594
10. 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 (Oxford, England) 26(8), 1112–1118 (apr 2010), http://www.ncbi.nlm.nih.gov/pubmed/20200009http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2853691
11. Norlin, L., Fransson, M.N., Eriksson, M., Merino-Martinez, R., Anderberg, M., Kurtovic, S., Litton, J.E.: A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. Biopreservation and biobanking 10(4), 343–8 (aug 2012), http://www.ncbi.nlm.nih.gov/pubmed/24849882
12. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., Mundlos, S.: The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American journal of human genetics 83(5), 610–5 (nov 2008), http://www.ncbi.nlm.nih.gov/pubmed/18950739http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2668030
13. Roos, M., Gray, A.J.G., Waagmeester, A., Thompson, M., Kaliyaperumal, R., van der Horst, E., Mons, B., Wilkinson, M.: Bring Your Own Data Workshops: A Mechanism to Aid Data Owners to Comply with Linked Data Best Practices. In: Paschke, A., Burger, A., Romano, P., Marshall, M.S., Splendiani, A. (eds.) Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences, Berlin, Germany, December 9-11, 2014. CEUR Workshop Proceedings, vol. 1320. CEUR-WS.org (2014), http://ceur-ws.org/Vol-1320/paper_36.pdf
14. Roos, M., Lopes, P.: Bring your own data parties and beyond: make your data linkable to speed up rare disease research. In: Vittozzi, L., Salvatore, M., Taruscio, D. (eds.) Abstracts presented to the EPIRARE International Workshop 24-25 November 2014. pp. 21–24. Rome (2014), http://rarejournal.org/rarejournal/article/viewFile/69/93
15. Samadian, S., McManus, B., Wilkinson, M.D.: Extending and encoding existing biological terminologies and datasets for use in the reasoned semantic web. Journal of biomedical semantics 3(1), 6 (jul 2012), http://www.ncbi.nlm.nih.gov/pubmed/22818710
16. Spackman, K.: Snomed rt and snomedct. promise of an international clinical terminology. M.D. computing : computers in medical practice 17(6), 29, http://www.ncbi.nlm.nih.gov/pubmed/11189756
17. Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., Robinson, P.N., Parkinson, H., Rath, A.: Ordo: An ontology connecting rare disease, epidemiology and genetic data. In: Phenotype data at ISMB2014 (2014), http://phenoday2014.bio-lark.org/pdf/9.pdf
18. Wolstencroft, K., Owen, S., Krebs, O., Nguyen, Q., Stanford, N.J., Golebiewski, M., Weidemann, A., Bittkowski, M., An, L., Shockley, D., Snoep, J.L., Mueller, W., Goble, C.: Seek: a systems biology data and model management platform. BMC Systems Biology 9(1), 33 (dec 2015), http://www.biomedcentral.com/1752-0509/9/33