IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

Meeting report series

Report of the 3rd DSC WG on Ontologies and Disease Prioritization teleconference

21 July 2014

Organization

Organized by: Scientific Secretariat
Teleconference

Participants

Dr Peter Robinson, Berlin, Germany (chair)
Prof Michael Bamshad, Seattle, USA
Dr Kym Boycott, Ottawa, Canada
Prof Melissa Haendel, Portland, USA
Prof Ada Hamosh, Baltimore, USA
Ms Janine Lewis, Bethesda, USA
Ms Suzanna Lewis, Berkeley, USA
Dr Chris Mungall, Berkeley, USA
Dr Yaffa Rubinstein, Bethesda, USA
Prof Pak-Chung Sham, Hong Kong, China
Dr Damian Smedley, Cambridge, UK
Ms Sharon Terry, Washington DC, USA

Dr Barbara Cagniard, Scientific Secretariat
Dr Sophie Höhn, Scientific Secretariat
Dr Lilian Lau, Scientific Secretariat

Apologies

Dr Michael Brudno, Toronto, Canada
Dr Helen Firth, Cambridge, UK
Ms Henrietta Hyatt-Knorr, Bethesda, USA
Dr Helen Parkinson, Cambridge, UK
Dr Ana Rath, Paris, France
Dr Maria Sobrido, Santiago de Compostela, Spain
Prof Maria Taboada, Santiago de Compostela, Spain

Agenda

- Recommendations for disease and phenotype ontologies
- Phenotype Exchange standards
- Biobanks and registries ontologies

REPORT

Each participant of the WG teleconference introduced herself/himself, as new members joined the Working Group. Participants have all a specific interest in ontologies and/or disease prioritization, either in the academic or industrial sector, with a majority of them working in the field of rare diseases.

Recommendations for disease and phenotype ontologies

The members of the Executive Committee agreed during their meeting in Berlin in May on the creation of a seal of approval named "Recommended by IRDiRC". This label will highlight tools and standards generated through IRDiRC activities, as well as tools and standards not generated by but identified by IRDiRC as key resources.

Process for adoption was defined:

- Scientific Committee to produce 1-2 pages of rationale, including list of other most important tools or standards, and contentious issues
- Consultation of the two other Scientific Committees
- Submission to the Executive Committee for approval

The Scientific Secretariat will provide further information about this process by e-mail.

Disease ontologies

List of ontologies

- ORDO (Orphanet Rare Disease Ontology)
- ICD (International Classification of Diseases)
- SNOMED (International Health Terminology Standards Development Organization)
- OMIM (Online Mendelian Inheritance in Man)

Orphanet is collaborating with EBI to create a formal rare disease ontology and to align it with the Disease Ontology which is more focused on complex common diseases.

Recommendation

Both ORDO and OMIM have complementary sets of qualities and advantages. The recommendation would be for IRDiRC to support both of them and to state that interoperability between both resources should be strengthened.

Phenotype ontologies

List of ontologies

- HPO (Human Phenotype Ontology)
- PhenoDB
- Orphanet
- POSSUM (Pictures Of Standard Syndromes and Undiagnosed Malformations; commercial ontology)
- SNOMED (commercial ontology, low coverage)

Recommendation

Both HPO and PhenoDB have complementary sets of qualities and advantages. PhenoDB is adequate for people looking for a smaller ontology than HPO. The recommendation would be for IRDiRC to support HPO and PhenoDB, and to state that these two resources should be mapped to one another, which they almost are.

The International Consortium for Human Phenotype Terminologies (ICHPT) was created in order to sort out the thousands of phenotypic terms appearing in the literature into a common vocabulary. Its mapping will be completed by the end of July. It is in line with the objectives of IRDiRC.

SNOMED and POSSUM are commercial ontologies, and SNOMED does not cover a lot of diseases.
Another recommendation would be that SNOMED needs to incorporate ICHPT terms to be more useful. The UMLS (Unified Medical Language System) recently analyzed the incorporation of HPO terms into SNOMED and showed that SNOMED has only about 30% of HPO terminologies. Moreover it is not easy to use for computational analysis, nor free, nor linked to many resources, not to mention the lack control over a commercial ontology which priority wouldn't necessarily aligned to IRDiRC's goals. IRDiRC should not recommend SNOMED as a primary ontology for rare diseases.
However, it was brought to attention that SNOMED is used by ClinGen, a National Institute of Health (NIH)-funded resource dedicated to harnessing both research data and the data from the hundreds of thousands of clinical genetics tests being performed each year, as well as supporting expert curation to determine which variants are most relevant to patient care. This project has decided to use SNOMED because the terms they needed were not present in the other ontologies. HPO and PhenoDB do not have a good coverage of cancers and metabolic diseases.

Phenotype Exchange standards

Standards

The Undiagnosed Disease Program (UDP), in collaboration with Michael Brudno, aims to create a data exchange format standard for phenotype data. A standardized syntax for sharing phenotypes data across different platforms such as UDP, the Monarch Initiative, and PhenomeCentral was needed. A text was drafted describing why none of the existing clinical data sharing standards did not match the needs for rare undiagnosed diseases, and what sort of technologies and approaches would be needed to help

diseases discovery for undiagnosed patients. This document aims to get community collaboration. It is a call for action to help define the standards in a technical way.

It would be important to have IRDiRC make a statement about Phenotype Exchange Standards. It would be ideal if the rare disease community could converge upon one standard that would be both flexible enough to allow people who are using different ontologies to use the same standard, and yet strict enough so that the software would be able to flexibly take requests from all over the world. One of the main use cases would be the Matchmaker Exchange group, which is linked to the ICHPT.

A very interoperable approach is required for phenotype data sharing. Interoperable and sophisticated description of the patients' phenotyping data will help matchmaking, and also help translational research to take advantage on undiagnosed diseases.

IRDiRC should recommend the Phenotype Exchange Standards as well as technical standards which include different phenotype ontologies and terminologies. Software would have to follow these standards so shared syntax and elements are used and encourages sharing between databases.

Matchmaking

Several efforts exist for matchmaking, amongst them notably mendeliangenomics.org, PhenomeCentral which was developed with UDP, DECIPHER (DatabasE of genomiC varIants and Phenotype in Humans using Ensembl Resources), and DDD (Deciphering Development Disorders). Most of them are open-source software. All of them have the ability to matchmake with inner databases.

Currently, an API (application programming interface) is being built by the Matchmaker Exchange group to allow data from each of these software, in theory, to be queried for matches. It is important to be able to work with other resources. For instance, the exchange data between PhenomeCentral and GeneMatcher is currently being tested. A draft paper is also currently being written on this API.

Ideally, to ensure a standard API is being developed, the active stakeholders (including IRDiRC and Global Alliance for Genomics and Health) should all convene for a meeting. Global Alliance group is planning a few small meetings during the ASHG in San Diego, the matchmaker group is likely to have a workshop. There will also be general workshop too. Members of this WG who are attending ASHG could represent IRDiRC in these workshops.

Mapping between phenotype ontologies

Phenotype Exchange Standards do not map the different phenotype ontologies. This is still an issue to be resolved. IRDiRC could make a statement about this issue, explaining that work and time are still needed.

Financial support

A workshop is planned by the NIH and NCI (National Cancer Institute) in November 2014 on phenotype data sharing for clinically actionable variants for both cancers and undiagnosed diseases. It might be possible that after this workshop, funding will be available to support the Phenotype Exchange Standards, at least in the USA from the NIH.

Biobanks and registries

Registry ontologies

A registry ontology would greatly help the Global Rare Diseases Patient Registry and Data Repository (GRDR^{SM}). A more standardized terminology is currently developed in the NIH which is composed of a set of core common data elements (CDEs). CDEs are defined by the core elements that any registry can ask (e.g. name of the patient, age, gender, ethnicity). These are not specific but general elements, considered core. A database where all the CDEs will be posted is currently under construction at the NIH and should be available in a month or so. The URL will be shared with the community when it becomes available.

In building a registry for the National Clinical Research Network, PCORnet have also improved a set of standardized CDE and coded them to provide a dictionary to interested parties. For rare diseases in PCORnet, there is also an attempt to look for a way to do computable phenotype across the clinical bodies in the US. PCORnet have also been urged to adopt principles discussed here, and it comes back to the point that a basic set of standards is important.

IRDiRC could recommend core CDEs. A list of registries standards could also be realized. A call should be realized in order to collect the standards people are using for their registries.

Standardization is important in the rare disease field. The GUID (Global Unique Identifier), assigned to the data of the patient, aims to facilitate the following of the data across studies, countries, registries and so on. To generate this GUID, there are certain elements which must be collected, such as the core common data elements. This is one example of standardization across registries.

IRDiRC should recommend the GUID.

Biobanks ontologies

Biobanks ontologies is the final area this WG will take care. This topic will be addressed during the next teleconference.

Main deliverables

- Send information on the label “Recommended by IRDiRC”
- Write the recommendations on disease and phenotype ontologies for distribution to all members for early August
- Comment and feedback on ontologies recommendation by the end of August
- Send the call for action paper on Phenotype Exchange standards to all members
- Send information about API exchange
- Write about the Phenotype Exchange Standards, the API and the mapping between phenotype ontologies
- Send the list of standards registries available on the GRDR website to the WG members – *Already done*
- Send information about the Biospecimen semantic workshop happening at the International Conference for Biomedical Ontologies
- Send information about the GUID and CDEs – *Already done*
- Contact people in the area of the GUID and CDEs to schedule the next teleconference
- Schedule the next teleconference for September/October