WP4 - Enabling CNV data discovery in diagnostic and phenotypic context

Project : ELIXIR hCNV community Implementation Study

Authors: David LAGORCE, Marc HANAUER

Reviews: David SALGADO, Michael BAUDIS, Christophe BEROUD

Version : 1

D4.1 Deliver a list of select ontologies required to efficiently capture phenotypic description useful for data interpretation for any genetic disease.

Ontologies are designed in order to represent the formal conceptualization of a specific domain. To this end, biomedical ontologies and controlled clinical terminologies are key in knowledge representation, data integration and interoperability. Based on a set of axioms, defined or constrained classes, they allow computational uses of the data in a meaningful way in several contexts including drug discovery, machine-learning technologies, natural language processing and widely ensuring decision support for health information systems and medical research. These kinds of ontologies are a means to capture and integrate research data across domains and over different levels of granularity. However, together with the proliferation of heterogeneous healthcare data, many biomedical fields are covered by numerous existing ontologies, disparate in quality and maintenance capabilities (Geller *et al* AMIA Annu Symp Proc 2018) resulting in the need of developing consensus recommendations for their appropriate selection within a biomedical community. Biomedical ontologies formally specify the meaning of terms in a vocabulary so that the meaning of the data can be understood and processed both by humans and machines. To express this meaning, ontologies utilize formal languages, as usual OWL, OBO or RDF. Species and/or domain-specific ontologies are now well developed and are available for human, mouse, fly, worm and yeast, for instance the "Mouse Developmental Anatomy Ontology". According to work published on the development of cross-species phenotype ontologies and their benefits, to date ontologies describing phenotypes exist for many species including e.g.the Mammalian Phenotype Ontology (**MP**), for Ascomycetes (**APO**) or for C. elegans (**WPO**). There are also well-established ontology design patterns for modeling phenotypes in a species and domain independent manner that utilize the Phenotype and Trait Ontology

(**PATO**) standard where a phenotype is characterized by an affected Entity (from an anatomy or process ontology) and a quality that specifies how the entity is affected.

Regarding phenotypic abnormalities in human disease, the Human Phenotype Ontology (**HPO)** is the current standard and an IRDIRC (International Rare Diseases Research Consortium) recognized resource as well as a part of the Monarch Initiative. **HPO** is a central component of one of the 13 driver projects in the Global Alliance for Genomics and Health (GA4GH) strategic roadmap and currently consists of 11,787 classes that provide a standardized vocabulary for describing phenotypic abnormalities encountered in human monogenic diseases. For human diseases aspects, other options include the Human Disease Ontology (**DO**) and the Orphanet Rare Disease Ontology (**ORDO**), available on the Elixir Core resource portal Orphadata.org. **DO** provides a classification of human diseases according to multiple axes related to genetic disorders, infectious diseases, metabolic disorders, cellular proliferations and others. It consists of 9247 classes that aim at unifying the representation of human diseases defined across a variety of developed biomedical vocabularies. **ORDO** provides a structured vocabulary for rare diseases and consists of 12,960 classes which provides a structured vocabulary to represent relationships between phenomes, diseases, genes and relevant features such genetic inheritance, supporting the computational analysis of rare diseases data. Further, Orphanet recently developed **HOOM** which is the ontological module linking rare disease to their phenotypes (by using **HPO** and **ORDO**).

Regarding ontologies at a cellular level, the Cellular Microscopy Phenotype Ontology (**CMPO**) provides a species-neutral controlled vocabulary for describing phenotypic qualities relating to the whole cell, cellular components, cellular processes and cell populations. Terms from CMPO are being used to annotate phenotype descriptions from high-content screening databases and cellular image repositories. Concerning genes, the Gene Ontology Consortium (**GO**) produces a dynamic, controlled vocabulary. One can retrieve in **GO** three independently rooted ontologies with specific purposes: *biological process* (refers to a biological objective to which the gene or gene product contributes), *molecular function* (defined as the biochemical activity (including specific binding to ligands or structures) of a gene product) and *cellular component* (referring to the place in the cell where a gene product is active).

For cancer related diagnoses and phenotypes, the **NCIt OBO Edition project** (https://www.ebi.ac.uk/ols/ontologies/ncit) has emerged as a well-structured and dynamically

developed resource which aims to increase integration of NCIt concepts with OBO Library ontologies. NCIt is a reference terminology that includes broad coverage of the cancer domain, including cancer related diseases, findings and abnormalities and is supported by projects translating such as "ICDOntologies" (https://github.com/progenetix/ICDOntologies) which provide mappings from and to classifications such as ICD-O.Apart from their use for phenotypes and diagnoses, ontologies are applied to the annotation of features and attributes of altered biological sequences for in oncology and rare disease contexts. Here, **SO** (Sequence Ontology) is dedicated to describe types of sequence alterations as well as associated biological process and features like binding sites and exons. SO also provides a rich set of attributes to describe these features such as “polycistronic” and “maternally imprinted”. Finally, **SEPIO** (The Scientific Evidence and Provenance Information Ontology) is capable of describing change in sequences like mutations, consequences of this mutation and the elements needed in order to categorize this mutation. This ontology supports description of evidence and provenance information for scientific claims. The core model represents the relationships between claims, their evidence lines, the information items that comprise these lines of evidence, and the methods, tools, and agents involved in the creation of these entities.

So, here is a list of exhaustive biomedical ontologies capable to capture phenotypic information:

Ontology Name	Exploitation range	URL
MP	Mammalian Phenotypes	http://www.informatics.jax.org/vocab/mp_ontology
PATO	Phenotypic qualities	http://www.obofoundry.org/ontology/pato.html
HPO	Human Phenotypes	https://hpo.jax.org/app/
ORDO	Human Rare Diseases	http://www.orphadata.org/cgi-bin/index.php#ontologies
HOOM	Rare Diseases / HPO	http://www.orphadata.org/cgi-bin/index.php#ontologies
GO	Human Genes	http://geneontology.org/

NCIt OBO	Oncology	https://www.ebi.ac.uk/ols/ontologies/ncit
SO	Biological sequences	http://www.sequenceontology.org/
SEPIO	Scientific claims	https://github.com/monarch-initiative/SEPIO-ontology

D4.2 Deliver lists of common data elements that should be provided in various situations such as rare diseases, oncology or common diseases

Copy Number Variation (CNV) is - by the extent of genome involvement - the most extensive form of genetic mutation, but gaps remain in our ability to share these mutation data and their associated phenotypes.

Within the ELIXIR hCNV community, the INSERM-AMU partner has previously developed **BANCCO** database for the **Achropuce** French Network of diagnostic laboratories that contains more than 31,696 CNV from 20,634 patients, CNV are stored using their genomic position, length, type (deletion, duplication) and their method of identification (cghArray, NGS, etc ...) using 3 reference genome coordinates (hg18, hg19 and GRCh38). Other CNV resources are built by partners in this community such as the **CIBERER CNV database** (ELIXIR-ES) and **Progenetix** (ELIXIR-CH). As the largest public resource for cancer CNV data the Progenetix database provides an overview of more than 500 types of human malignancies. It combines genome CNV profiles of currently more than 115'000 cancer samples with disease concepts using NCIt, UBERON and ICD-O classifications, and also provides ontology mapping and API services. All these CNV repositories are already or in progress to be interfaced with Beacon protocols.

Besides these resources developed by members of the hCNV community, various other databases connected to the ELIXIR ecosystem provide CNV data within varying general scopes such as Gnomad SV repository, Ensembl, Decipher and VKGL.

Common Data Elements (**CDE**) are standardized key terms or concepts, established so that they may be used in clinical research or in studies, to enhance data quality and so that the data can be used across sites and over time. The **CDE** is similar to an attribute; it functions as a key, which can then map to an associated value. Development and use of **CDE** supports standardization of terms and facilitates data sharing so that data can be compared and combined across studies; research findings can then be generalized with respect to different research institutions, diverse populations, different regions, and interventions. In this sense, NIH define **CDE** as an element that is common to multiple data sets across different studies with a certain type: *Universal* regardless of condition of interest*, Domain-specific* for use in studies of a particular topic and Required as a matter of institutional policy.

CDE can be applied to describe either genomic impact and phenotypic description of a mutation.

Phentotypic and link to the disease

According to the specificity of Rare Diseases a list of standards exists which are an important building block of the European Platform on Rare Disease Registration (EU RD Platform) and works on recommendations to further increase the FAIRness of CDEs in RD registries (https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en). Released as the first practical instrument for increasing interoperability of RD registries, it endorses 16 data elements perceived to be essential for RD research, thus should ideally be employed by all RD registries across Europe. It provides a major cornerstone for the field. Towards meeting FAIR principles, it recommends using a number of other standards to denote CDEs, including ORPHAcodes, HGVS-nomenclature (Human Genome Variation Society), HGNC (Hugo Gene Nomenclature Committee), OMIM (Online Mendelian Inheritance in Man), and ICF (International Classification of Functioning and Disability). Also, the EUPID Services provide mechanisms for patient re-identification, if needed. Initially, the EUPID Services were developed in the Pediatric Oncology community with its adoption now being extended to, and presently implemented, for the rare disease community as a whole. Regarding linkage between rare diseases and phenotype ontologies, these have been a core RD standardization topic for many years. Widely used, effective solutions now exist, not least the Human Phenotype Ontology (**HPO**), the Orphanet Rare Disease Ontology (**ORDO**). Further, Orphanet has developed the ontological module linking rare disease to their phenotypes (**HOOM**). Thus, HOOM is able to link these both previous ontologies and then is highly recommended for CNV description. Recently, as part of the Phenotypic Data Exchange work supported by the Global Alliance for Genomics and Health (GA4GH), the PhenoPackets standard (https://phenopackets-schema.readthedocs.io) (ISO/ WD 4454) has been developed to enable the exchange of clinical phenotype related information between information systems. Its scope includes data on individuals, biological samples, sequence variants as well as and family/pedigree information. A core concept in PXF is the general preference for data objects referring to established ontologies (e.g. HPO, ORDO, NCIt, OMIM, and HOOM in the future.) using compact identifiers (CURIES).

Description of the mutation

Regarding "Omics" data standards and file formats, GA4GH is now the umbrella organization providing widespread standards like SAM/BAM (Sequence Alignment Map/Binary Alignment Map) and CRAM (Compressed Reference-oriented Alignment Map) for sequence alignments, VCF/BCF (Variant Call Format/Binary Call Format) for variant listing, and crypt4GH for storage of genomic data in an encrypted and authenticated state. Additional mature genomic standards in this area include BED, GFF, FASTA and FASTQ.

GA4GH variant annotation (https://ga4gh-gks.github.io/variant_annotation.html) and variant representation (https://ga4gh-gks.github.io/variant_representation.html) task forces are working on defining formalisms to describe a mutation with ambiguity. SV and CNV are planned to be included in this formalisms. The GA4GH HTS-Spec working group (http://samtools.github.io/hts-specs/), including ELIXIR hCNV partners, is working on a new VCF specification v4.4 (http://samtools.github.io/hts-specs/VCFv4.4.draft.pdf), to allow the description of CNV and more broadly Structural variations using this reference file format.
Mature standards also exist for other omics data types. Transcriptomic information consists of positional intensity/signal and feature expression information. Positional data are managed using the bed/bigbed/bigwig standards, while feature data in the form of a matrix employ a tab delimited format or the LOOM file format (based on HDF5).

Other omics file formats

Genomics file formats, such as SAM/BAM, BED or FASTQ, can also be used for transcriptomics data. For proteomics, the ProteomeXchange Consortium (http://www.proteomexchange.org) provides data storage options for mass spectrometry (MS) using recognized standards such as mzML (an open, XML-based format for mass spectrometer output files) and mzIdentML (an open, XML-based data standard for peptide and protein identifications). Relevant here are ISA-Tab or ISA-JSON (https://isa-tools.org/format/specification.html) as multi-omics data exchange formats, along with the Omics Markup Language (OML, ISO/DIS 21393). Finally, it also exists Domain-agnostic Interoperability Standards that are designed for specific domains or purposes (e.g., regarding distinct data types such as genes and/or symptoms). In this context we could cite the Data Catalogue Vocabulary (DCAT), RDF/JSON-LD, Web Ontology Language (OWL), the Linked Data Platform, JSON-related «formats, the Digital Object Interface Protocol (DOIP), and its “FAIR” derivative the FAIR digital object. Furthermore, for some use cases (not least automated machine reasoning), we will benefit from creating semantic models for other standards using the W3C web technology stack, so that they are interoperable at the level of the generic data model provided by the resource description framework (RDF). Also, relevant here are Schema.org and Bioschemas. Schema.org is a specification to mark-up data resources on the web with basic semantic terms to improve findability by search engines such as Google. It describes types of information (items which can be described), which then have properties (the descriptions). The Bioschemas community (sponsored by ELIXIR) proposes new types and properties (e.g. for

BioChemEntity or Taxon) to Schema.org, including types such as genes, molecular entity or diseases, profile which is in development phase.