WSBO-2021

Workshop on Synergizing Biomedical Ontologies

Program & Abstracts

July 14 – 16, 2021
Virtual

Stephan Schürer[1], University of Miami
Mark Musen[1], Stanford University

Samantha Jeschonek[2], Collaborative Drug Discovery
Asiyah Yu Lin[2], National Institutes of Health
Hande Küçük McGinty[2], University of Ohio

[1] WSBO-2021 Chair
[2] WSBO-2021 Organizer

Program – WSBO 2021

Day 1: Wednesday, July 14

all times are in EDT

Moderators: Stephan Schürer and Samantha Jeschonek

8:00am – 8:20am	**Welcome and Opening Remarks** *Stephan Schürer*
8:20am – 8:40am	**Synergizing Biomedical Ontologies - An Industry Perspective** *Martin Romacker*
8:40am – 9:00am	**Ontology Mappings - Untangling the Hairball and Democratising the Results** *David Osumi-Sutherland*
9:00am – 9:20am	**Matching Life Sciences Ontologies in the Ontology Alignment Evaluation Initiative** *Ernesto Jiménez-Ruiz*
9:20am – 9:40am	**FAIRifying Biomedical Ontology Synergy** *Lynn Schriml*
9:40am – 10:00am	**Break**
10:00am – 10:40am	Keynote Presentation: **The Dangers of Bad Mappings: How Imprecise and Incomplete Mappings Can Cost Lives** *Melissa Haendel*
10:40am – 11:00am	**Towards Automating the Ontological Representation of Proteins in the Protein Ontology** *Darren Natale*
11:00am – 11:20am	**A Simple Standard for Sharing Ontology Mappings (SSSOM)** *Nicolas Matentzoglu*
11:20am – 11:50am	**Breakout Session: Needs for Synergization in the Biomedical Ontology Community** *Moderator: Stephan Schürer*

Day 2: Thursday, July 15

all times are in EDT

Moderators: Mark Musen and John Turner

12:00pm – 12:20pm	**Opening Remarks** *Mark Musen*
12:20pm – 1:00pm	Keynote Presentation: **FAIRness and Fairness in Sharing Data from Study Participants and Patients** *Lucila Ohno-Machado*
1:00pm – 1:20pm	**Bridging the Phenotype Divide by Using Shared Patterns** *Nicole Vasilevsky*
1:20pm – 1:40pm	**Synergizing Biomedical Ontologies with Genomics Databases** *Christopher Mungall*
1:40pm – 2:00pm	**OntoloBridge: Connecting Ontology Users and Ontology Maintainers** *John Turner*
2:00pm – 2:20pm	**Break**
2:20pm – 2:40pm	**Resolving Ontology Mappings using Boomer** *James Balhoff*
2:40pm – 3:00pm	**Development and Maintenance of the Interoperable and Synergistic Cell Line Ontology** *Oliver He*
3:00pm – 3:20pm	**What End Users Need from Ontology Community? - Experience from NCPI, FDA, and COVID-19 Ontologies Harmonization Effort** *Asiyah Lin*
3:20pm – 3:40pm	**CDD Annotator and Perspectives from the Data FAIRy Initiative** *Samantha Jeschonek*
3:40pm – 4:10pm	**Breakout Session: Facilitating Interoperability via Software** *Moderator: Asiyah Lin*

Day 3: Friday, July 16

all times are in EDT

Moderator: Hande Küçük McGinty

11:00am – 11:10am	**Opening Remarks** *Hande Küçük McGinty*
11:10am – 11:25am	**Automating Ontology Mapping Workflows With ROBOT** *Rebecca Jackson*
11:25am – 11:30am	**Harmonizing Units of Measure Vocabularies on the Web: a Basic Prototype** *Kai Blumberg*
11:30am – 11:40am	**Thinking about the Future: Ontologies for Whole-Person and Holistic-Healthcare Research** *Hande Küçük McGinty*
11:40am – 11:50am	**OBO Foundry Domain Coverage for Food, One Health, and Holistic Healthcare** *Damion Dooley*
11:50am – 12:20pm	**Describing the Need: An Ontology End-User Case Study** *Sheryl Denker and Antal Berényi*
12:20pm – 12:30pm	**Collaborative Drug Discovery – Supporting the Future of FAIR Data Management** *Whitney Smith*
12:30pm – 12:50pm	**Closing Remarks** *Stephan Schürer and Mark Musen*
12:50pm – 1:10pm	**Conference Feedback** *Samantha Jeschonek*

Abstracts

Day 1: July 14, 8:20am – 8:40am EDT

Synergizing Biomedical Ontologies – An Industry Perspective

Martin Romacker[1]

Working with semantic technologies and advocating for the broad adoption of ontologies has been pretty exotic in the Pharmaceutical Industry during the past decades. After this long period of hibernation in pharma companies, the situation has completely changed. The key driver of this fundamental change is the digital transformation of the industry promoting the insights, analytics and data business. The FAIR data principles have emerged as a common theme across all pharma to underpin the digital transformation and to ensure that data are managed in a way to make them findable, accessible, interoperable and reusable.

Considering the FAIR Maturity indicators, semantic technologies, RDF-based community vocabularies, formal standards and ontologies outline the major ingredients for Data FAIRification at scale. Consequently, the internalization of ontologies and their correct integration in our data management value chain have become core activities of Information Architects and Data Engineers. The biomedical domain benefits from a wealth of semantic resources which is both a blessing and a curse. On the one hand, we benefit from the sheer number of existing resources, on the other hand there is a lot of redundancies (many GUPRIs for the same entity) as well as diverging modeling principles and quality issues hampering interoperability and mapping.

In the presentation, we would like to give a glimpse on the work we have been doing at Roche to bring semantic resources into our data ecosystems. But we will also highlight related challenges when working with biomedical ontologies. We would like to promote the idea that synergizing ontologies should result in clear community business rules for ontology development but also convergence meaning that less resources would be more.

[1] *Data and Information Architect, Roche Innovation Center in Basel*

Day 1: July 14, 8:40am – 9:00am EDT

Ontology Mappings - Untangling the Hairball and Democratising the Results

David Osumi-Sutherland[1]

Nicolas Matentzoglu[2], James McLaughlin[3], Henriette Harmse[3], Susan Bello[4], Nicole Vasilevsky[5], James Balhoff[6], Christopher Mungall[7], Melissa Haendel[8], Helen Parkinson[3]

Mapping between ontologies can be essential to data integration, but ontology mappings have long been second-class citizens in the ontology world - relegated to simple links within ontology files or shared mapping tables lacking formalization. The all too frequent result is a confusing tangle of often conflicting mappings with no record of provenance or indication of quality. The quality of mappings is being improved by two approaches.

We are involved in major efforts to harmonise ontology semantics via the use of common relations, shared classes and term templates. This semantic convergence is driving improvements in semantic similarity-driven mappings between ontologies. For example, the uPheno project uses this approach to align the logical definitions across phenotype ontologies for model organisms and humans, resulting in improved accuracy of semantic similarity-based mapping used in variant prioritization.

New approaches are emerging to untangle the mapping hairballs that result from the independent assertion of pairwise mappings between multiple ontologies in a domain. This issue is especially prominent in the disease ontology world, where we have been working to resolve this in the Mondo disease ontology using a combination of Bayesian ontology merging - combining probabilistic and logical inference - coupled with manual curation to fix resulting inconsistencies.

Improvements to mappings mean little without better ways to formalise and share them. We have developed a Simple Standard for Sharing Ontology Mappings (SSSOM) that records mapping metadata including provenance, mapping type (lexical, logical, equivalent, broad, narrow) and confidence. As well as facilitating FAIR sharing of mappings, SSSOM underpins improvements to the EBI Ontology Xref Service (OxO), allowing users to map identifiers between their ontologies via its web app or API leveraging the highest quality mappings available. In this talk I will illustrate this using examples from uPheno and Mondo.

[1] *Ontology Developer, EMBL/EBL*, [2] Semanticly, [3] EMBL/EBI, [4]The Jackson Laboratory, [5] Oregon Health & Science University, [6] Rnaissance Computing Institute (RENCI), [7] Lawrence Berkely National Laboratory, [8] University of Colorado

Day 1: July 14, 9:00am – 9:20am EDT

Matching Life Sciences Ontologies in the Ontology Alignment Evaluation Initiative

Ernesto Jiménez-Ruiz[1]
Thomas Liener[2], Ian Harrow[3]

The Ontology Alignment Evaluation Initiative (OAEI) is an annual campaign for the systematic evaluation of ontology matching systems. The main objective is the comparison of ontology matching systems on the same basis and to enable the reproducibility of the results. The OAEI 2020 included 12 different tracks, 4 of which involved life sciences ontologies. The Pistoia Alliance has been actively involved in the OAEI co-organising the Disease and Phenotype track. Systems participating in the OAEI have the potential of being applied in real-world scenarios in combination with human curation. Some ontology matching systems, like AML and LogMap, are not only able to find correspondences across ontologies, but they can also point to logic-based compatibility issues when merging the relevant ontologies with a candidate set of mappings.

[1] *Lecturer in AI, University of London,* [2] Pistoia Alliance Inc, [3] Ian Harrow Consulting

Day 1: July 14, 9:20am – 9:40am EDT

FAIRifying Biomedical Ontology Synergy

Lynn M. Schriml[1]

Clarifying the boundaries between ontological domains to minimize overlap, addresses domain fuzziness and recognizes usage differences thus enhancing synergy across ontologies. Within ontological domains, the OBO Foundry's "Scope" principle addresses the ideal of "non-overlapping and strictly-scoped content". Expanded usage of a domain through adaptations or derivatives are "part in parcel of" the evolution of biomedical knowledge. Our challenge here is to 're-synergize' ontological terms that have drifted apart, in order to FAIRify biomedical resources utilizing ontologies to standardize and harmonize datasets. Currently, the FAIRness of biomedical ontology synergy is significantly challenged by scope fuzziness across ontologies within domains. While the ideal for "biomedical ontology synergy" is to 'reuse' the term ID, label and definition from a source ontology, the "reality" is that cross references (xrefs) are often utilized to indicate 'EquivalentTo' mappings within biomedical domains. This practice has resulted in a significant loss of synergy, term proliferation, decreased FAIRness of associated datasets and the loss of interoperability. Realistically, use case driven development of ontologies includes a mishmash of 'ideal' and 'situationally appropriate' approaches. Within the disease domain, for example, even the original definition of "disease" (DOID:4 in the Human Disease Ontology) has been 'reused' word for word or slightly reworded and assigned to either new IDs or renamed (disease or disorder). To address these synergy challenges we need a new approach for FAIRifying biomedical ontologies within domains. Recognizing the synergy challenges, emphasizing synergy best practices and openly discussing novel synergy approaches that respect individual development can move synergy efforts forward. Defining the provenance of 'reference ontology' IDs will enhance "Reusability" of ontology terms and biomedical data. For example, one solution for improving Findability and Accessibility could be to include 'reference ontology IRIs' (via an "imported from" Annotation Property), rather than creating xrefs for 'source biomedical ontologies'.

[1] *Associate Professor, University of Maryland School of Medicine, Institute for Genome Sciences*

Day 1: July 14, 10:00am – 10:40am EDT

KEYNOTE PRESENTATION

The dangers of bad mappings: How imprecise and incomplete mappings can cost lives

Melissa Haendel[1]

Diseases are represented in different terminologies, for different demographics, regions, contexts, and in different databases. Many resources map across systems, but these mappings often lack provenance or specifics, and do not often take into account attributes of a disease such as phenotypic, genetic, and environmental characteristics. Further, different databases model the attributes of a disease and their relationships differently. For example, one resource may include the variant-to-disease association, whereas another records only the phenotypic features and their onset associated with the disease. As knowledge advances, there is also a philosophical debate about what constitutes a disease, and when to "lump" or "split". For example, if two diseases share the same phenotypic profile but have variations in different genes, are they the same disease? Further, individual variation in disease attributes often does not align with reference definitions of diseases. This collective lack of agreement negatively impacts diagnosis and treatment while complicating research on mechanisms and therapies. There is an urgent need to precisely and globally define rare diseases - and their mappings - to save lives.

[1] *Chief Research Informatics Officer, University of Colorado Anschutz Medical Campus*

Day 1: July 14, 10:40am – 11:00am EDT

Towards Automating the Ontological Representation of Proteins in the Protein Ontology

Darren Natale[1]

PRO is the OBO Foundry ontology that provides an ontological representation of protein-related entities. Each PRO term represents a distinct class of entities at various levels of abstraction, including evolutionarily-related protein families, orthologous gene products, orthologous isoforms, specific proteoforms, and protein complexes, ranging from the taxon-neutral to the taxon-specific. Considering the large number of known proteins, we automate import whenever possible. For example, taxon-specific canonical gene product and isoform terms can be automatically created based on UniProtKB entries, while specific proteoforms can be generated by import of information within Reactome or the Top-Down Proteomics resource. As well, we have developed—and plan to enhance—a resource to generate PRO terms dynamically upon user request. Terms created by import from one resource are cross-referenced back to the source entries and, when possible, to identical entries from another resource. Here, we describe these pipelines and the challenges encountered.

[1] *Research Assistant Professor, Protein Information Resource at Georgetown University Medical Center*

Day 1: July 14, 11:00am – 11:20am EDT

A Simple Standard for Sharing Ontology Mappings (SSSOM)

Nicolas Matentzoglu[1]

James P. Balhoff, Tiffany Callahan, Christopher Chute, Jiao Dahzi, William Duncn, Davera Gabriel, John Graybeal, Melissa Haendel, Henriette Harmse, Solbrig Harold, Nomi Harris, Harshad Hedge, Charles Tapley Hoyt, Ernesto Jimenez-Ruiz, Simon Jupp, Hyeongsik Kim, Sebastian Koehler, Thomas Liener, James Malone, James McLaughlin, Monica Munoz-Torres, David Osumi-Sutherland, James Overton, Anne Thessen, Nicole Vasilevsky, Chris Mungall

Term mappings are of fundamental importance to interoperability, yet often lack metadata to be correctly interpreted and applied in contexts such as data integration or transformation. For example, are two terms equivalent or merely associated? Are they narrow or broad matches? etc. Such relationships between the mapped terms often remain unclear, which makes them very hard to use in scenarios that require a high degree of precision (such as diagnostics or risk prediction). Furthermore, the lack of metadata on the methods and rules involved in producing the mappings and confidence estimations regarding their correctness makes it hard to combine and reconcile mappings, especially curated and automated ones.

Working as part of a large collaborative group, we have developed a Simple Standard for Sharing Ontology Mappings (SSSOM) which addresses these problems by 1.introducing a simple vocabulary for mapping metadata 2.defining an easy to use table-based format that can be integrated into regular data science pipelines without the need to parse or queryontologies 3.defining a set of exports formats such as RDF/XML and JSON-LD and SQL tables.

SSSOM is defined using a LinkML schema (https://linkml.github.io), and defines metadata for many key features of term mappings and mapping sets, such as mapping confidence, versioning, mapping tools and match types (lexical, logical, human-curated). We show how we use SSSOM tables to curate, combine and disseminate mappings between phenotype, disease and anatomy ontologies, provide an overview of our Python toolkit for processing mappings in SSSOM format and sketch out our plans to more finely represent mapping curation rules for clinical and other use cases. The working draft of the SSSOM specification can be found at http://w3id.org/sssom/SSSOM.md

[1] *Independent Semantic Technology Consultant, Semanticly Ltd.*

Day 2: July 15, 12:20pm – 1:00pm EDT

KEYNOTE PRESENTATION

FAIRness and fairness in sharing data from study participants and patients

Lucila Ohno-Machado[1]

In our quest for harmonization of data collected from clinical studies or electronic health records, we discuss very specific issues in biomedical ontologies, (meta)data standards, etc. This technical discussion is critical for the effective use of shared data. However, the perspective of those whose data being shared should not be forgotten. In this presentation, I will discuss the reasons why we need data from human participants or patients to be FAIR (findable, accessible, interoperable and reusable) but also fair (used with participants' permission, representing the diversity of the population, allowing researchers from a range of institutions to have access).

[1] *Associate Dean for Informatics and Technology, UC San Diego Health, Department of Biomedical Informatics*

Day 2: July 15, 1:00pm – 1:20pm EDT

Bridging the Phenotype Divide by Using Shared Patterns

Nicole Vasilevsky[1]

Susan Bello[2], Nicolas Matentzoglu[3], David Osumi-Sutherland[4], The Upheno Team[5]

To facilitate the identification of useful models of human disease and support cross-species gene to phenotype relations we need to ensure interoperability between the various taxon-specific phenotype ontologies that are used to annotate these data. The use of equivalence axioms (EQs) within ontologies provides machine-readable, logical term definitions that when applied consistently help bridge the phenotype divide. These EQs define phenotypes using logical patterns that utilize cross-species and species-neutral ontologies to build definitions, including UBERON (anatomy), PATO (qualities), and GO (biological processes, cellular components, and molecular functions). However, building consistent and precise EQs across phenotype ontologies is a challenge. The uPheno project was formed to meet this challenge by aligning EQs across phenotype ontologies and developing a library of shared patterns. As part of this work, inconsistent patterns are identified and reconciled. For example, we discovered that the Mammalian Phenotype and Human Phenotype ontologies used two different patterns to logically define cystic phenotype terms (e.g. Hepatic cysts, HP:0001407; prostate gland cysts, MP:0009737). One pattern used the PATO term “cystic” (PATO:0001673) and the other used the MPATH term “cyst” (MPATH:62). After review, a common pattern for cyst phenotypes using the PATO term (cystInLocation.yaml) was developed and used to reconcile these terms. The shared patterns developed by the uPheno group provide a common framework for building EQs that may be used manually or in an automated fashion within the Ontology Development Kit (ODK). Use of the patterns within the ODK also allows for rapid revision and addition of terms using a common pattern, reducing the labor involved in pattern maintenance. The uPheno project GitHub site (https://github.com/obophenotype/upheno) provides access to the shared patterns and a forum to bring new issues forward for community discussion. Funded by NIH / NHGRI #1RM1HG010860

[1] *Visiting Associate Research Professor, University of Colorado*, [2] The Jackson Laboratory, [3] Semantically, [4] EMBL/EBI, [5] University of Colorado Denver

Day 2: July 15, 1:20pm – 1:40pm EDT

Synergizing Biomedical Ontologies with Genomics Databases

Christopher Mungall[1]

Many ontologies have terms that overlap with entities that are already represented in existing genomics databases or other major resources. For example genes, proteins, pathways, organismal taxa, chemical entities, drugs, cell lines. This can result in duplication of effort, and confusion the part of users trying to select the appropriate identifier to represent concepts of interest. Furthermore, it compounds already difficult mapping issues.

In this presentation I give an overview of some key genomics resources, and the ways in which existing ontologies create unnecessary and confusing duplication, and ways in which unintended logical incoherence can occur.

I then present our own work on creating isomorphic database to ontology conversions that are intended to work in concert with existing databases, avoiding creation of new mappings. These are in the area of (1) genes and other molecular entities that have evolved to serve a biological role; (2) chromosomes and chromosomal regions, in particular the monochrom ontology (3) organismal taxa derived from NCBITaxon; and (4) chemical entities. I also show how the pathway-process dichotomy is resolved in the context of the Gene Ontology, Reactome, and the Pathway Ontology.

Finally I will present a proposed scheme in which databases, knowledge graphs, and ontologies can co-exist, each playing to their individual strengths, without the need to generate de-novo identifiers and compound mapping issues.

[1] *Department Head, Biosystems Data Science Berkeley Lab*

Day 2: July 15, 1:40pm – 2:00pm EDT

OntoloBridge: Connecting Ontology Users and Ontology Maintainers

John Turner[1]

Formalized vocabularies are important tools to enable researchers to Find, Access, Integrate and Reuse (FAIR) scientific datasets and other digital research objects. Building and maintaining ontologies is an ongoing process that requires significant resources and collaboration from domain experts and ontologists. It therefore frequently happens that terms required for annotating a digital resource are not available in an ontology. This limits data curators and is a barrier for the wide-spead use of many ontologies. OntoloBridge is a FAIR-bridge technology to connect that gap between users of controlled scientific vocabularies and the creators of the underlying ontologies. OntoloBridge thus encourages and enables community input as a crucial feature in the widespread use of ontologies. We have developed OntoloBridge as a semi-automated technology along with a public API that allows users to request new terms and update existing ones. The OntoloBridge alpha release currently serves the following highly-accessed ontologies: BioAssay Ontology (BAO), Drug Target Ontology (DTO), Protein Ontology (PR), Cell Line Ontology (CLO), and the Coronavirus Infectious Disease Ontology (CIDO). As the system matures, we aim to incorporate many other ontologies. Easy access is provided via the BioPortal where registered users can request a “New Term” via a simple form available under the Classes section. OntoloBridge has also been fully integrated into the BioAssay Express (BAE) human-machine hybrid annotation system for bioassays. Via BAE, users can suggest ontology terms on the fly and initiate feedback between curators and ontology experts. Ontolobridge can also be accessed via the API and we are working on a standalone website. The Initial user reaction to OntoloBridge has been very positive, and has increased productivity by streamlining annotations. OntoloBridge can likely make an impact in fast pace projects such as COVID-19 research.

[1] *Research Associate, University of Miami*

Day 2: July 15, 2:20pm – 2:40pm EDT

Resolving Ontology Mappings Using Boomer

James P. Balhoff[1]
Christopher J Mungall[2]

When using mappings to combine existing ontologies into a cohesive whole, it can be a challenge to untangle the logical implications resulting from converting mappings into equivalence or subclass relations, especially where mappings connect multiple overlapping ontologies or vocabularies. However, creating a logical interpretation for a set of mappings is a valuable strategy for assessing the mutual coherence of a large body of interrelated mappings. We have created a tool, Boomer, which uses a combined logical and probabilistic approach to translate mappings into logical axioms that can be used to merge ontologies. Boomer implements a search algorithm to find the combined ontology with the highest probability that is also logically coherent. Boomer takes as input ontologies, plus SSSOM mappings with probabilities for OWL interpretations for each mapping and produces as output a set of cliques representing a particular combination of interpretations, together with a visualization of that clique. The resulting graphics provide a focused view into problematic tangles of mappings; this lends itself extremely well to an iterative inference process with a curator in the loop.

Boomer is implemented in Scala, and is built on the Whelk OWL reasoner. Whelk uses immutable data structures which allow Boomer to readily roll back to previous reasoning states in the course of its search algorithm. Boomer's exhaustive search tests axiom combinations in order of decreasing joint probability. For larger cliques Boomer runs many depth-first "greedy" searches, with different starting points performed in parallel.

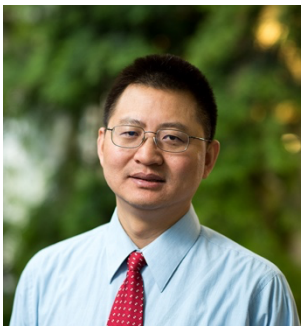
We are currently using Boomer to resolve mutual cross-references between Gene Ontology molecular functions, the Rhea reaction knowledgebase, Enzyme Commission entries, and MetaCyc pathways. Problematic clique images are posted as issues in the Gene Ontology GitHub repository for evaluation by GO editors. A previous implementation of Boomer (k-BOOM) was used to create the initial version of the Mondo disease ontology.

[1] *Research Scientist, Renaissance Computing Institute*, [2] Lawrence Berkeley National Laboratory

Day 2: July 15, 2:40pm – 3:00pm EDT

Development and maintenance of the interoperable and synergistic Cell Line Ontology

Yongqun Oliver He[1]



The Cell Line Ontology (CLO) is a community-based biomedical ontology in the domain of cell line cells. Since its initial publication in 2014, many changes have been made. For example, CLO has now covered thousands of stem cell line cells and terms associated with stem cell line investigation. CLO also includes cell line cells from China that are expressed in both English and Chinese. We have also systematically represented the cell line cells studied in the Library of Integrated Network-Based Cellular Signatures (LINCS) project. The basic design pattern has also been expanded to cover more areas. Currently CLO has included over 40,000 cell lines. CLO has been developed to become interoperable and synergistic with other ontologies in the OBO library. Tools such as OntoloBridge and GitHub have been used to support our interoperable ontology development and maintenance.

[1] *Associate Professor, University of Michigan Medical School*

Day 2: July 15, 3:00pm – 3:20pm EDT

What End Users Need from Ontology Community? - Experience from NCPI, FDA, and COVID-19 Ontologies Harmonization effort

Asiyah Lin[1]
Alex Welsch[2], Lawrence Callahan[2]

Multiple ontologies covering similar area exists, such as MONDO, HPO and DO for disease or phenotypes, which becomes a challenge for end users who are generally non-ontology developers. Use cases such as NIH Cloud Computing Interoperability (NCPI)'s dataset catalog dashboard, FDA Global Substance Registration System (GSRS) will be presented to ask ontology community for a solution on annotating conditions, indications with one reliable source that encompass MONDO, HPO, DO and MeSH. The requirement of such resource will be discussed. An example of ontology harmonization effort of COVID-19 ontologies will be presented to show a possible solution. Challenges and short coming for ontology harmonization solution will be discussed.

[1] *Data and Technology Advancement Scholar, NIH*, [2] FDA

Day 2: July 15, 3:20pm – 3:40pm EDT

CDD Annotator and Perspectives from the Data FAIRy Initiative

Samantha Jeschonek[1]

Isabella Feierberg[2], Whitney Smith[3], Jason Harris[3], Alex Clark[3], Tim Ikeda[4], Rama Balakrishnan[5], Martin Romacker[5], Dana Vaderwall[6], Anosha Siripala[7], Gabriel Backiananthan[7], Vladimir Makarov[8], Thomas Liener[8]

Collaborative Drug Discovery (CDD)'s Annotator (formerly, BioAssay Express (BAE)) provides an innovative solution for managing assay metadata. Powered by a patented hybrid machine learning model and leveraging expertly curated ontologies, assays can be annotated in less than five minutes with superior accuracy. Structuring assay metadata converts previously unstructured descriptions into machine-operable formats that comply with FAIR (findable, accessible, interoperable, and reusable) principles. Metadata annotation provides scientists with an additional axis of information for analysis, for the first time allowing them to harness the power of assay informatics.

In 2020, a team of scientists from AstraZeneca, Bristol Myers Squibb, Novartis, and Roche set forth to find a way to convert unstructured, public biological assay descriptions into FAIR information objects. This initiative, known as the DataFAIRy BioAssay Annotation Project, selected CDD's Annotator as the technology of choice due to its emphasis on accuracy, usability, and ability for dynamic and iterative human-feedback. The Data FAIRy pilot was completed in January 2021.

In this talk, we will present our Annotator technology along with the lessons learned in the pilot project to annotate bioassay descriptions *en masse* and will chart a way to expand this effort in the future.

[1] *Product Manager, Collaborative Drug Discovery*, [2] Jnana Therapeutics (previously at AstraZeneca), [3] Collaborative Drug Discovery, [4] AstraZeneca, [5] Roche, [6] Schrödinger (previously at BMS), [7] Novartis, [8] Pistoia Alliance

Day 3: July 16, 11:10am – 11:25am EDT

Automating Ontology Mapping Workflows with ROBOT

Rebecca C. Jackson[1]

The process of mapping an ontology to one or more other ontologies can be a tedious one. Developers must not only ensure that the mappings are high-quality and accurate, but they also need to implement these mappings within their ontology project. Once the mappings have been added, they must also be verified, adding yet another task to the workflow. We present the ROBOT tool for use in creating automated ontology mapping workflows, so that ontology developers can focus on the quality and accuracy of the mappings, rather than spending time on repetitive technical tasks.

ROBOT (a recursive acronym for "ROBOT is an OBO Tool") is an open-source Java library and command-line tool for automating ontology development tasks. Most usage is through the command-line tool, which runs on macOS, Linux, and Windows. ROBOT provides commands for a variety of broad ontology development tasks, including converting formats, running a reasoner, generating quality control reports, and more. More specifically, many of these commands can be used in workflows involving ontology mappings, such as extracting import modules and creating new axioms from spreadsheet templates. These commands can also be combined into larger workflows using a separate tool like GNU Make, allowing ontology developers to automate their ontology mapping processes.

In this presentation, we will go over some of these commands and how they can be used in larger automated workflows.

[1] *Ontologist and Software Developer, Bend Informatics LLC*

Day 3: July 16, 11:25am – 11:30am EDT

Harmonizing Units of Measure Vocabularies on the Web: an Initial Prototype

Kai Blumberg[1]

With increased focus on making data FAIR, the need for a unified system of units of measure is paramount. Today there are several unit systems developed for and used by various communities from medical, marine, food, and engineering disciplines. Here we present progress toward a harmonized vocabulary for units on the web, which maps and unifies existing standards within a semantic web framework.

[1] *Ontologist, Biological and Chemical Oceanography Data Management Office (BCO-DMO)*

Day 3: July 16, 11:30am – 11:50am EDT

Thinking about the Future: Ontologies for Whole-Person and Holistic-Healthcare Research

Hande Küçük McGinty[1]

As holistic approaches to healthcare research gain attention, we will discuss how food and agriculture ontologies are needed to complement bio-ontologies. We believe food and agriculture ontologies need to work concordantly with bio-ontologies to better explore multi-component interventions for the whole-person and holistic-healthcare research. We will discuss how evolving ontologies systematically and in a semi-automated way may help and accelerate the integration of knowledge across different domains.

OBO Foundry Domain Coverage for Food, One Health, and Holistic Healthcare

Damion Dooley[2]

Research and development projects are seeking standardized vocabulary across all direct food supply activities ranging from agricultural production, harvesting, preparation, food processing, marketing, distribution and consumption, as well as indirectly, within health, economic, food security and sustainability analysis and reporting tools. A cluster of food-related agriculture, farm animal health surveillance, diet, food metabolism, drug interaction, and nutritional study ontologies have recently joined OBO Foundry, thus spawning both a curation community and a vision of a greater integrated framework with which to study human health related to One Health and the farm-to-fork journey. A holistic view however would take OBO even further into the domains of consumer behaviour, socioeconomic indicators, planetary health, and even a cultural and philosophical re-examination of the human relationship to food.

[1] *Research Associate Professor, Ohio University*
[2] *Ontology Development Lead, Public Health Bioinformatics Group at Simon Fraser University*

Day 3: July 16, 11:50am – 12:20pm EDT

Describing the Need: An Ontology End-User Case Study

Sheryl Denker[1]
Antal Berenyi[2], Aimee Allen[2], Jennifer Drake[2], Ellen Berg[2]

Ontologies are essential in harmonizing and integrating disparate data for computational analyses including data processing, statistical analyses, and data mining. They are a key component of data governance, a discipline that ensures the validity and consistency of data assets. They are also central for optimal findability and consumption of commercial bioassays and services by the community of academic, industry and government scientists searching for tools to advance public health through development of new medicines. This broad community may be searching product databases, websites, product distributor sites, scientific or grey literature. We recognize a potential need for a uniform understanding of terms, data, and metadata such that a uniform understanding of biological terms exists among users. Two well-known, online distributors of biological assays and services, each with its own specific research area categories (ontology classes) and terms used for searching the sites, were evaluated for this case study. Are these terms broadly understood? Are these terms the same between the two sites? In an effort to assign similar terms related to biological assays to both online distributors, we took a combination manual and natural language processing (NLP) approach. This work in progress involves 1) researching a broad range of ontologies in order to find the most relevant, 2) establishing evaluation criteria to use in guiding a decision on choosing the most relevant ontology for Eurofins Discovery Services, and 3) practical application of similarity scoring in ontology mapping.

[1] *Senior Strategic Content Manager, Eurofins Discovery,* [2] Eurofins Discovery

Day 3: July 16, 12:20pm – 12:30pm EDT

Collaborative Drug Discovery – Supporting the Future of FAIR Data Management

Whitney Smith[1]
Samantha Jeschonek[2], Barry Bunin[2]

Easy – Secure - Collaborative. CDD has delivered secure, performant, and affordable research informatics solutions for over 15 years, earning the reputation as our industry's most trusted provider for cloud-based drug discovery data management. Following the tremendous success of CDD Vault, we have recently brought that same philosophy and expertise to the domain of assay metadata management and curation with CDD Annotator. In 2021, we have also leveraged these same principles to launch an exciting new platform, BioHarmony – a subscription data service that provides continually updated, annotated, and minable data on known drugs, gathered from across the semantic web and organized by a suite of proprietary technologies. This talk will explore these new platforms and our vision for FAIR data management in a drug discovery ecosystem that is increasingly scientifically diverse and organizationally collaborative.

[1] *Director of Business Development, Collaborative Drug Discovery,* [2] Collaborative Drug Discovery

WSBO 2021 Chairs

Stephan C. Schürer, Ph.D.

Associate Professor in the Department of Molecular and Cellular Pharmacology at the University of Miami and Program Director of Drug Discovery at the Center for Computational Science (CCS).

Dr. Schürer has over 15 years of research and management experience in industry and academia working on data standards, -integration, -modeling, scientific content and software development, cheminformatics and data science. Over the years, he has been involved in many small molecule probe and drug discovery projects and have been working in leadership positions in several national multi-site research consortia including the NIH Molecular Libraries Program, the Library of Integrated Network-based Cellular Signatures (LINCS), Big Data to Knowledge (BD2K), and the Illuminating the Druggable Genome (IDG) projects.

The Schürer research group applies distributed and parallelized bio- and chemoinformatics tools and builds modeling pipelines to understand drug mechanism of action, –specificity, –promiscuity and –polypharmacology. To physically make and test the most promising small molecules, the group develops computationally-optimized synthetic routes and applies parallel synthesis technologies to synthesize small compound libraries. Together these capabilities enable efficient prioritization of novel cancer-relevant drug target hypotheses, identification of precision drug combinations, and the development of novel small molecule tool compounds as potential starting points for drug discovery projects.

Mark A. Musen, M.D., Ph.D.

Professor of Biomedical Informatics and Biomedical Data Science at Stanford University and Director of the Stanford Center for Biomedical Informatics Research.

Dr. Musen conducts research related to open science, intelligent systems, computational ontologies, and biomedical decision support. His group developed Protégé, the world's most widely used technology for building and managing terminologies and ontologies. He served as principal investigator of the National Center for Biomedical Ontology, one of the original National Centers for Biomedical Computing created by the U.S. National Institutes of Health (NIH). He directs the Center for Expanded Data Annotation and Retrieval (CEDAR), founded under the NIH Big Data to Knowledge Initiative. CEDAR develops semantic technology to ease the authoring and management of biomedical experimental metadata. Dr. Musen directs the World Health Organization Collaborating Center for Classification, Terminology, and Standards at Stanford University, which has developed much of the information infrastructure for the authoring and management of the 11th edition of the International Classification of Diseases (ICD-11).

Dr. Musen was the recipient of the Donald A. B. Lindberg Award for Innovation in Informatics from the American Medical Informatics Association in 2006. He has been elected to the American College of Medical Informatics, the Association of American Physicians, the International Academy of Health Sciences Informatics, and the National Academy of Medicine.