Broad, Interdisciplinary Science *In Tela*: An Exposure and Child Health Ontology

James P. McCusker
Sabbir M. Rashid
Zhicheng Liang
Yue Liu
Katherine Chastain
Paulo Pinheiro
Rensselaer Polytechnic Institute
Troy, NY, USA

Jeanette A. Stingone
Icahn School of Medicine at Mount Sinai
New York, NY, USA

Deborah L. McGuinness
Rensselaer Polytechnic Institute

ABSTRACT

Data curation for interdisciplinary collaborative science requires a new online web-based approach that integrates domain knowledge from multiple resources and enables *in tela* (in the web) interactive collaboration between data providers, domain specialists, and data analysts. The Children's Health Exposure Analysis Resource (CHEAR) is a resource for child development and environmental exposure data. The CHEAR Data Center has developed an ontology that integrates study and exposure data in a way that is consistent across the program, and integrates with many best practice relevant vocabularies and repository schemas. This includes the World Wide Web Consortium's recommended Provenance Ontology (PROV), Semanticscience Integrated Ontology (SIO), the Chemical Entities of Biological Interest (CheBI) ontology, the Uberon multi-species anatomy ontology, and the Units Ontology as the starting point for our domain modeling. We mapped terms where they overlapped and extended these ontologies with classes that were required to support modeling and integrating data from epidemiology and chemical exposure measurements that comprise the majority of the data recorded by the CHEAR data center. In response to this challenge, we used an on-demand approach to develop the ontology based on a set of representative pilot projects in CHEAR. After initial development, we evaluated the ontology for completeness in representing an additional pilot study. An epidemiologist was able to produce a mapping of the project to the ontology with only minor corrections needed by an ontology expert. In the large dataset that was tested, one third of the classes needed to represent the dataset needed to be added to the ontology, all of them in areas where we expected to see more ontology expansion. Our overall approach is to drive towards completion of coverage while being relatively easy to use for domain experts. Ultimately we aim to have domain experts handle the majority of extensions and evolution with small interactions with ontology experts. In this paper, we report on our on-demand approach for web-based collaborative interdisciplinary ontology development and maintenance and also introduce the resulting extensible and interoperable exposure and child health ontology.

1 INTRODUCTION

Interdisciplinary science typically requires collaboration across multiple laboratories, often involving people with a wide variety of expertise areas, and it is critical to align data across multiple organizations typically in multiple domains. When these efforts evolve into large scale collaborations, as in the case of the National Institute of Environmental Health Sciences (NIEHS)-funded Children's Health Exposure Analysis Resource (CHEAR), the challenges also increase. In order to support the alignment, we are developing an interdisciplinary ontology addressing issues related to exposure and health at a variety of levels of granularity and from a wide range of perspectives. Challenges arise when developing such an interdisciplinary ontology relating to determining what should be modeled, at what level of detail, what vocabularies, schemas, and ontologies should be used as starting points, what alignments are needed, and what gaps need to be addressed. To manage the representational load and great diversity in content, we have attempted to apply the YAGNI (You Aren't Gonna Need It) principle from software engineering [11], only adding something on demand when we are sure we will need it. One question is how does YAGNI mesh with managing appropriate modeling of diverse data collections resulting from epidemiologically motivated studies? The data that the CHEAR program needs to manage covers several domains, and the data gathered by epidemiologists, while containing some commonalities, is diverse in its consideration of social, environmental, developmental, educational, and other factors, making the potential space of data representation large. These studies are themselves a combination of medical, social, behavioral, and economic data with environmental exposure data, that in turn, covers a broad set of chemical analysis, proteomic, genomic, and epigenomic data.

CHEAR supplements existing epidemiological studies with exposure data in an effort to provide a more holistic setting for data analysis over individual and pooled data. One goal is to provide a resource that includes an integrated representation of many independently-funded studies across the public health domain. In this broad interdisciplinary setting, as in most situations, it was essential to

WebSci'17, June 25-28, 2017, Troy, NY, USA.

978-1-4503-4896-6/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3091478.3091497

pick a core set of ontologies to provide a starting point. We looked for well used, well designed ontologies that were extensible, that covered key areas, and that were maintained by appropriate groups on which to base our work. In this paper, we report on the very positive experience we had using on-demand approach to ontology generation and maintenance in combination with web standards and enhanced web-based semantic infrastructure. We found that we can facilitate efficient collaboration by supporting and integrating to a common view of data from epidemiology, materials analysis labs, and -omics (proteomics, genomics, and others) specialists to provide a coherent view for data analysts. This paper provides insight into our approach, describes some of the benefits, and also introduces the resulting interdisciplinary and extensible exposure and health ontology.

2 RELATED WORK

While there is a range of related work on which to build, no single ontology and in fact, no set of ontologies adequately covers the breadth and depth needed for exposure science, health, and development as a foundation for integrative data repositories. Only one ontology that we know of provides explicit support for environmental exposures, the Environmental Exposures Ontology (ExO) [18]. It focuses on modeling exposure events, but does not itself support data representation. Other ontologies and knowledge graphs, such as ChEBI [4], PubChem [29], and UniProt [13] provide identifiers for specific chemicals and changes to blood chemistry, but do not model how they relate to outcomes or to conventional epidemiological variables.

2.1 Ontology Development

Much of the current literature on ontology development involves the application of either top-down or bottom-up approaches, specifically the ones reviewed in [14] and [3]. Our approach, called the "Semantic eScience Methodology" [8], initially developed for virtual observatories, proposed a middle out approach, which was use case driven and iterative, and grows the representations into specificity without focusing initially on alignment with an upper level ontology. We extend this approach through two main directions. First, we have developed additional tools to more tightly integrate domain experts into the ongoing ontology development team through use of tabular editing tools. Second, we focus on coverage of data representation for a growing set of studies. We use these studies as use cases to determine requirements and priorities for expansion. Third, we start with a core set of ontologies that provide a common upper- and mid-level ontology representation in appropriate areas. These provide a lens through which to organize our data representation efforts.

2.2 Foundational Ontologies

In order to leverage the existing best practices work from the ontology and relevant domain communities, the CHEAR Ontology builds on several foundational ontologies. The ontologies discussed here are all integrated and organized in the Human-Aware Science Ontology (HAScO) [24], which the CHEAR Ontology in turn imports. The Semanticscience Integrated Ontology (SIO) [5] defines types and relations for objects, processes and attributes, and therefore provides the integrated framework from which the ontology is rooted. SIO is a self-contained ontology that provides support for information resources; as well as hypothetical, fictional, and imaginary entities, and it has seen increasing usage, particularly in biomedical settings. It is not integrated with other ontologies, but we were able to easily integrate it with other ontologies in HAScO. In order to capture the provenance of concepts included in the ontology, such as the source or how a term was generated, we use the World Wide Web Consortium's (W3C)'s [16] recommended language for provenance on the web - the PROV-O ontology. In order to maintain details about data acquisition measurements, such as instruments or analytical methods used, the Virtual Solar Terrestrial Observatory Instruments (VSTOi) Ontology [8] and the Human-Aware Sensor Network Ontology (HASNetO) are leveraged [24]. The Units Ontology (UO) is used to incorporate a taxonomy of both English and Metric units into the ontology [9]. While UO does not cover every possible unit of measure, we have not yet had a need to represent units that go beyond the ones in UO. As with all ontology integration efforts, there were some mismatches in modeling perspectives. For example, while SIO treats units as OWL individuals, UO treats them as classes. We solved this mismatch using punning (in OWL 2) the UO classes to individuals as they are used. For terms related to chemical entities and anatomy, we turn to the Chemical Entities of Biological Interest (CheBI) [4] and Uberon multi-species anatomy [23] ontologies, respectively. CHeBI contains a large fraction of the chemicals analyzed in CHEAR, but not yet all of them. CHeBI's entities are well-defined with useful data about chemical structure, biological roles, and other details. We replicated the CHeBI style for the entities that we needed to add and also submitted extension requests to the CHeBI team for future integration, but in the meantime, those terms are in our CHEAR ontology. We also use PubChem [15] as a source of additional chemical entities that aren't yet in CheBI. Uberon contains all of the biospecimen types that we expect to be processing in CHEAR, except for buccal cells. The disease ontology (DO) [27] and ExO provide good starting points for capturing disease and exposure terms. While DO is not as comprehensive as ICD, it is focused on a more scientific treatment of diseases, instead of classifying diseases for medical administrative purposes, which has been the primary use case of ICD. Our statistical classes, especially variable types (like z-score) and methods are borrowed on demand from the Statistics Ontology (STATO).[1] Finally, we use properties from the Simple Knowledge Organization System (SKOS) [22] and The Dublin Core Terms (DC-Terms) [12] vocabularies to annotate classes in our ontology.

3 DEVELOPMENT METHODS

We developed the CHEAR ontology by extending the Human-Aware Science Ontology (HAScO)[2] (which in turn integrates the PROV-O and SIO ontologies). HAScO provides a standardized base for many of our scientific data-oriented projects, and serves as the integration point for our core science ontologies. HAScO contains concepts for representing various studies, as well as the data acquisition process,

[1] http://stato-ontology.org/

[2] http://hadatac.org/ont/hasco/

Table 1: Table of ontologies used in the CHEAR Ontology organized by role. Imported ontologies are adopted wholesale into CHEAR, while Annotation ontologies are used for concept metadata. MIREOT (Minimum Information to Reference and External Ontology Term) ontologies have been identified as on-demand sources for ontology extension. These ontologies are compatible with CHEAR-O, and some of their classes have been included in CHEAR-O using MIREOT principles [2].

Role	Ontology	Prefix	URI
Imported	SIO [5]	sio	http://semanticscience.org/resource/
	PROV-O [16]	prov	http://www.w3.org/ns/prov#
	Units Ontology [9]	uo	http://purl.obolibrary.org/obo/UO_
	HAScO [24]	hasco	http://hadatac.org/ont/hasco#
	HASNetO [24]	hasneto	http://hadatac.org/ont/hasneto#
	VSTO-I [8]	vstoi	http://hadatac.org/ont/vstoi#
Annotation	SKOS [22]	skos	http://www.w3.org/2004/02/skos/core#
	DC Terms [12]	dc	http://purl.org/dc/terms/
MIREOT	CheBI [4]	chebi	http://purl.obolibrary.org/obo/CHEBI_
	STATO	stato	http://purl.obolibrary.org/obo/STATO_
	PubChem [15]	pubchem	http://rdf.ncbi.nlm.nih.gov/pubchem/compound/
	Uberon [23]	uberon	http://purl.obolibrary.org/obo/UBERON_
	Disease Ontology	do	http://purl.obolibrary.org/obo/DOID_

including instruments, deployments and platforms involved. From a web semantics standpoint, HAScO provides a schema for modeling the acquisition process and attributes, which in turn allows for mapping of study variables to ontology concepts. This gives us our foundation on which we base our ontology work. We then map terms to each other as needed and identify ontology term gaps that must be filled by reviewing the CHEAR data specifications and the CHEAR study data dictionaries, codebook terms and laboratory information. We use LabKey [3], a web-based Laboratory Information Management System (LIMS), to gather and curate class definitions, annotations, and hierarchies. LabKey is also used to manage identifiers for subjects and samples across CHEAR studies. LabKey supports creation of "lists", essentially spreadsheets, that we use to generate the ontology from a Semantic Extract, Transform, and Load (SETL) process that is described for SETLr.[4] Both domain experts and ontology engineers have been using the web-based LabKey lists as the means to collaborate on ontology development by contributing and reviewing ontology definitions. SETLr is a tool for transforming data from tabular, Extensible Markup Language (XML), and JavaScript Object Notation (JSON) data sources into Resource Description Framework (RDF). This approach of building the ontology from class definitions curated in a LIMS allows domain experts to provide content for and review the ontology within a tabular format, organized through tables of things like analyte types, sample types, roles, and epidemiological attributes. It does this by using ETL-like workflows and templating based on existing web and semantic standards, like JSON-Linked Data (JSON-LD) [1], PROV-O [16], and python libraries like Pandas [21], Jinja2[5], and RDFlib.[6]

The overall process is detailed in Figure 1. For CHEAR, we primarily focus on the representational needs of the Pilot Study data dictionaries and codebooks, as well as the data reporting templates created by the CHEAR program. For each potential class we identify, we first check the ontology and foundational ontologies to see if there is an adequate class. If there is an exact match, we note it in a mapping table. If there is a partial match, we subclass that match with the needed specialization. New classes are added to one of several LabKey lists, depending on the subtree it is part of. We have lists in LabKey for subtypes of object, attribute, role, sample, analyte, and process. More complex class definitions that are not yet supported by the SETLr conversion process are included as ontology fragments in Turtle. The SETL script enumerates the fragments and LabKey lists and processes them all into a single RDF graph. The classes and their definitions, labels, and alternate labels are reviewed by domain experts, and the lists are processed by SETLr to create a generated ontology. This ontology is imported into HADatAC, where curators and reviewers can further comment on it. That feedback is incorporated into LabKey as well.

The resulting ontology is published to the web. [7] This initial ontology was developed using five pilot studies that were determined to be representative of the studies that are anticipated to be submitted to the CHEAR program. The five pilot studies consisted of four prospective birth cohorts and a clinical study of pediatric allergic disease. In order to represent the expected future studies that would enroll in CHEAR, the pilot studies were selected to cover multiple health outcomes and multiple critical windows of exposure and development including pregnancy, early infancy and childhood. The birth cohorts included questionnaire and biological data on mother-child pairs while the clinical study also included treatment data extracted from the child's medical record. One of the birth cohorts was conducted within an international study population. While all pilot studies had stored biospecimens that would be analyzed within the CHEAR laboratories, the biological matrices (sample types) varied by study and included serum, urine and placenta.

With each study, we added classes that were determined to be necessary to represent the data as presented in those studies. As

[3] http://www.labkey.com/

[4] https://github.com/tetherless-world/setlr/

[5] http://jinja.pocoo.org

[6] http://rdflib.readthedocs.io

[7] http://hadatatac.org/ont/chear/

we accept more studies, the CHEAR ontology will be expanded to support those studies. We anticipate a long-tail function in the need to create additional classes, once we have covered the most common classes. This long tail is consistent with the YAGNI principle - the most commonly needed capabilities are introduced at first, and as new capabilities are needed, that are added to the system (or ontology, in our case).

We have initial evidence that this approach is working well. We used one of the initial pilot studies to evaluate coverage for the required terms. We had excellent coverage except in two areas (fertility methods and BMI rating categories), where we had not attempted to provide any modeling primitives, so this gap was anticipated. Based on the evaluation, we had 74% coverage of demographics, 100% coverage of environmental exposures, and 57% of health/disease outcomes. Given the small number of initial pilot studies and the anticipated gap filling requirements for diseases as related content emerges, we were impressed with these statistics. As we reviewed one additional birth-cohort pilot we found similar results. We had approximately 67% total coverage.The primary gaps were in investigator-defined categorizations of existing concepts, like education.

We found the ontology useful to anyone who needs to model data in epidemiology, especially with studies that perform significant sample analysis. It is ready to be used as a public health data interchange standard, and we expect that studies that use the ontology will find most of the classes they need in it. We are actively recruiting users of the ontology and are encouraging requests if the ontology is missing terminology needed to cover study data in the general areas of exposure and child health.

We did not need to add any new properties to the CHEAR ontology, as the approach of SIO is to create new attributes as objects rather than statements. This also allows us to add information about when the attribute was determined (*sio:measuredAt*), if the attribute was measured in relation to something else (*sio:inRelationTo*), what the unit of measure is (*sio:hasUnit*), (for instance, concentration of a substance in relation to the sample the substance is from), as well as provenance information, like what other measurements (or other objects) an attribute is derived from (*prov:wasDerivedFrom*) and what methods were used (*prov:wasGeneratedBy*).

In Figures 2, 3, 4, and 5 we show how fundamental types of data are represented. Figure 2 shows how child, mother, study, and household relate in a set of core social/familial relationships, using subclasses of *sio:Role* and the *sio:hasRole*. Figure 3 shows how basic measurements are expressed via subclasses of *sio:TimeInterval*. Anthropometry attributes, like head circumference, are expressed on objects that are a type of head (from the UBERON Ontology) that have an attribute of *sio:Circumference*. The head objects are then stated to be *sio:isPartOf* the subject the measurement was taken from. Figure 4 shows how laboratory analysis can be expressed, by deriving a sample from the subject, and expressing the sio:Concentration (*sio:inRelationTo* the sample) of a particular molecular entity. Most of these entities are identified in either ChEBI or in PubChem. Figure 5 shows how the most common genomic, transcriptomic, and epigenomic data can be represented. The figure covers Single Nucleotide Polymorphisms (SNPs) and Variations (SNVs), genomic region Copy Number Variations (CNVs), gene expression, and regulatory site methylation fraction.

The ontology was developed initially using the CMap Ontology Edition (COE) tool to model coverage of the initial four pilot studies in the CHEAR program with a data scientist/ontologist and an epidemiologist. This was aggregated into a set of lists stored in LabKey, broken out by subtrees in the ontology. These lists continue to be curated by data scientists, epidemiologists, and ontologists to extend and improve the ontology as needed. The official ontology is generated using the Semantic Extract, Transform, and Load-r (SETLr) by writing an ontological interpretation of the lists and their columns. A number of classes, like Z-Fenton Birth Weight [7], BMI, and others have been enriched with extended OWL restrictions that are difficult to express in a tabular format. We include those definitions in OWL/RDF fragments that are combined during the SETL process. The resulting ontology is published (with incremented versioning) at http://hadatac.org/ont/chear/, and is used by the HADatAc (Human Aware Data Acquisition Framework) system as a basis for ingesting data from the CHEAR program for search, discovery, and curation.

While our approach does not attempt a one to one mapping of ontology concepts to datasets, we attempt to provide the ability to compose descriptions of specific measurements, like "Mother's Pre-Pregnancy BMI" from concepts contained in the ontology. A dataset could be annotated with a formal definition that uses OWL property restrictions that *compose* into the attribute of interest. For example, the OWL Manchester notation [10] for "Mother's Pre-Pregnancy BMI" (here encoded as a column called "MPPBMI") would be:

```
Class: MPPBMI
SubClassOf: chear:BMI,
  sio:measuredAt only chear:PrePregnancy,
  sio:isAttributeOf only (
    sio:hasRole some (
     chear:Mother and sio:inRelationTo only chear:Child
    )
  )
```

Similarly, *chear:BMI* is defined in terms of the attributes it uses, and records the required unit of measure:

```
class: chear:BMI
SubClassOf: sio:Quantity,
  prov:wasDerivedFrom some sio:Height,
  prov:wasDerivedFrom some sio:Mass,
  sio:hasUnit value uo:000086 # kg/mˆ2
```

This allows us to be much more flexible in how we map data from a given study into the CHEAR Ontology. Concentrating on mapping many data sets to one single conceptual structure serves the semantic web goal of interoperability: any dataset for which the mapping is completed has a mapping to the SIO conceptualization, and can be compared to any other dataset that has also been mapped.

4 EVALUATION

We attempted to annotate a new pilot study with the CHEAR ontology as a way to determine what the coverage of the current version of the ontology was. Our epidemiologist created an initial semantic version of a data dictionary with our lead ontology expert.

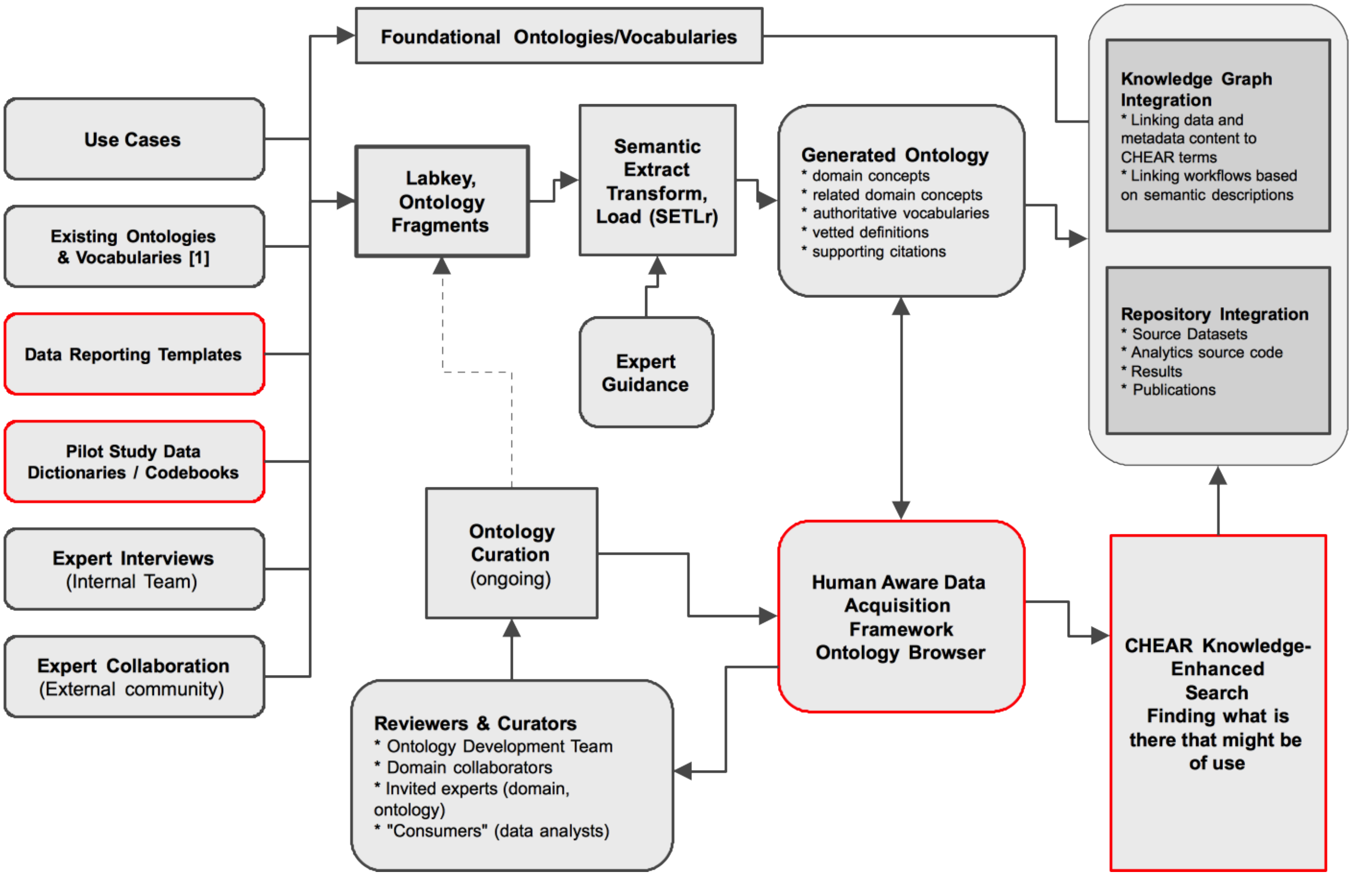
Foundational Ontologies/Vocabularies
Use Cases
Existing Ontologies & Vocabularies [1]
Data Reporting Templates
Pilot Study Data Dictionaries / Codebooks
Expert Interviews (Internal Team)
Expert Collaboration (External community)
Labkey, Ontology Fragments
Semantic Extract Transform, Load (SETLr)
Expert Guidance
Generated Ontology
* domain concepts
* related domain concepts
* authoritative vocabularies
* vetted definitions
* supporting citations
Knowledge Graph Integration
* Linking data and metadata content to CHEAR terms
* Linking workflows based on semantic descriptions
Repository Integration
* Source Datasets
* Analytics source code
* Results
* Publications
Ontology Curation (ongoing)
Human Aware Data Acquisition Framework Ontology Browser
CHEAR Knowledge-Enhanced Search Finding what is there that might be of use
Reviewers & Curators
* Ontology Development Team
* Domain collaborators
* Invited experts (domain, ontology)
* "Consumers" (data analysts)

Figure 1: Ontology Development process for the CHEAR Ontology.

The epidemiologist indicated where classes were missing from the ontology and wrote suggested concepts and definitions.

The annotated data dictionary was then curated by an ontology expert to correct any mistakes, which were minimal, and to find existing terms in our adopted ontologies only minor corrections needed by an ontology expert. Of the 96 classes used to represent the data, we needed to introduce 32 new classes, 16 of which were in the study codebook, mapping enumerated values to attributes. Additionally, of the 32 new classes, 11 of them mapped into existing classes in the set of MIREOT ontologies.

We find this level of class introduction is acceptable for early stage ontology development. The new classes will be released in the CHEAR Ontology version 0.9,[8] while the version of the ontology used for analysis is version 0.8.[9]

5 DISCUSSION

The YAGNI principle has been a guiding force for prioritizing capabilities in software. As the name "You aren't gonna need it" implies, this principle recommends incorporation of only essential functionality. While sometimes it can result in technical debt due to a lack of foresight, when properly managed and paired with a long term vision of how a work will grow, can result in powerful and relevant software. YAGNI can sometimes result in technical debt. However, vision does not need to immediately lead to implementation. Maintaining a long-term vision of how the software or ontology will grow while only implementing what is immediately needed helps keep scope and, as the vision's elements become relevant, also validates that vision.

Similarly, we found that when it was applied to ontology development, YAGNI can quickly find the most relevant classes to model, and when paired with a useful set of core ontologies, can create a monotonic development process, minimizing the need to change existing classes. This approach of on-demand ontology extension has proven to us to be a scalable approach, since, as new data comes in, fewer and fewer classes are needed. New classes have fit easily within existing ones. One component with significant variability is the mention and encoding of education level as granaurlity varies rather significantly and education paths also vary in different countries, so not surprisingly, education level has been encoded quite differently in some datasets. We did find that we could leverage the general monotonicity of education - one generally completes previous grade levels before proceeding to more advanced grades. There were some challenges around aligning education levels and terminologies across countries. We found the

[8] http://hadatac.org/ont/chear/0.9/
[9] http://hadatac.org/ont/chear/0.8/

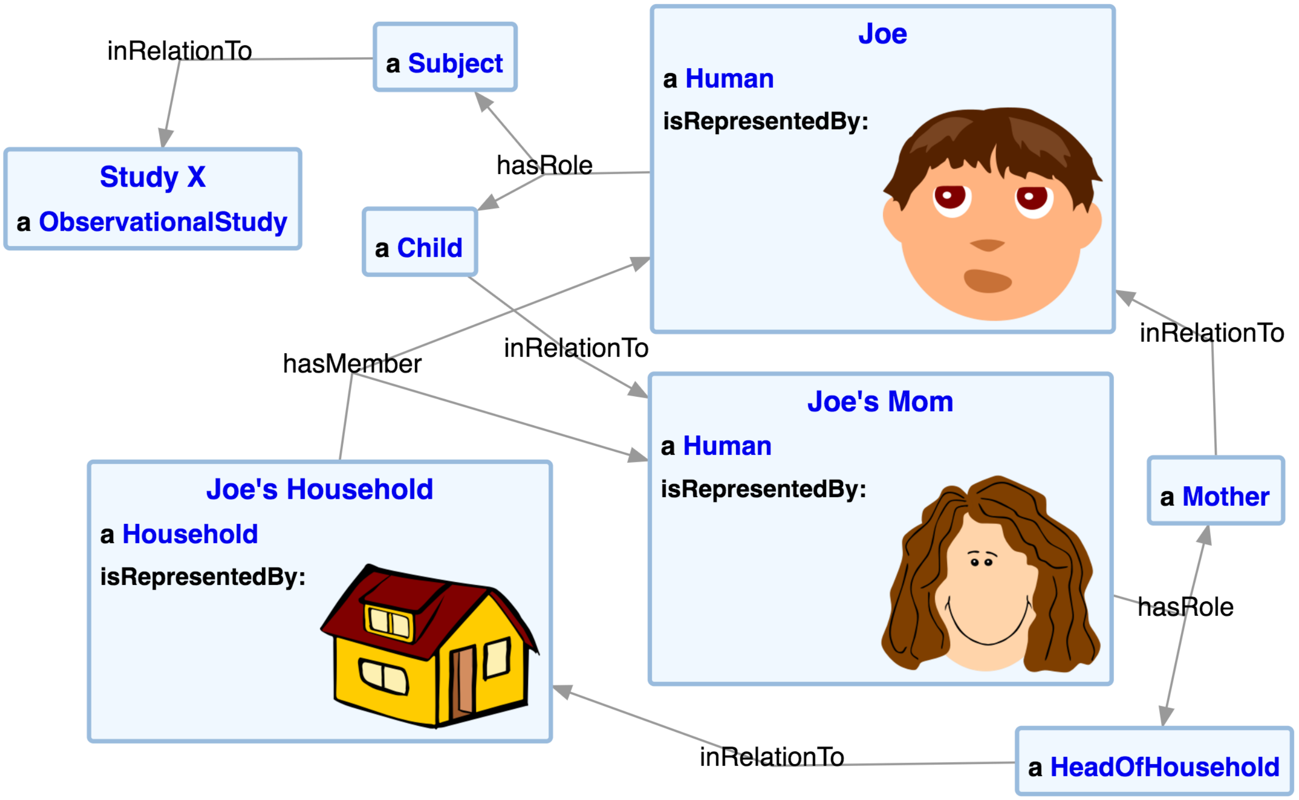

Figure 2: Representation of Familial Roles and Relations. Roles (Child, Subject, Mother, Head of Household) link entities (Joe, Joe's Mom, Joe's Household) together via *sio:hasRole* and *sio:inRelationTo*. Household membership is indicated via *sio:hasMember*.

combination of PROV-O, SIO, HAScO, and HASNetO to be a very useful core ontology set for describing scientific data. In fact, we found it to provide almost surprisingly good coverage for the types of studies we have seen so far. Coverage of entities, roles, attributes, samples, lab analysis variables, instruments, methods, and even -omics data was straightforward to model.

5.1 Representation Considerations

There are a number of representations that might be common to some data representation-driven ontologies, but are not used in the CHEAR Ontology. The term ”observation” has been used with a range of meanings and it can be ambiguous and sometimes misleading. Because we record attributes as reified objects, they mostly serve in the role that observation objects might, including time of measurement, provenance of the attribute, unit of measure, and other attributes. This approach is consistent with the use of ontologies that attempt to model the world as described, as opposed to language-oriented ontologies, that attempt to model the description of the world.

We subscribe, for the purposes of this ontology, to Platonic realism, where classes are allowed to be made that predict unobserved, hypothetical instances in the world, or fictional instances that are marked as such. This form of scientific realism, which SIO was designed for, was laid out in [6]. We therefore borrow, at times from ontologies that, in their details, align well with SIO, even if their upper level ontologies do not follow quite the same philosophical underpinnings.

One important consideration to make when using this approach is to carefully pick the foundational ontologies that are used. When dealing with ontologies that may have different approaches to representation, it may be necessary to select subtrees from an ontology to use. If the ontology outright prohibits that approach by including statements that lead to reasoning contradictions, it may be necessary to take relevant subtrees, and, where allowed by license, duplicate them with references to the source. For instance, if a conceptual ontology, like SNOMED-CT [28], LOINC [20], or MeSH [17] has a useful, consistent subtree that can be used in a realist ontology, a 1:1 mapping can be made using SPARQL queries into an ontology module. Each class can be related back to the originating URI using a property from the Conceptual Model Ontology (CMO) called *hasPrimaryConcept* [19]. CMO provides a generalized framework for associating realist ontologies that focus on data models with conceptual ontologies that are often used for coding or similar applications. Attribution can also be used by stating that the copy *prov:wasDerivedFrom* the original version.

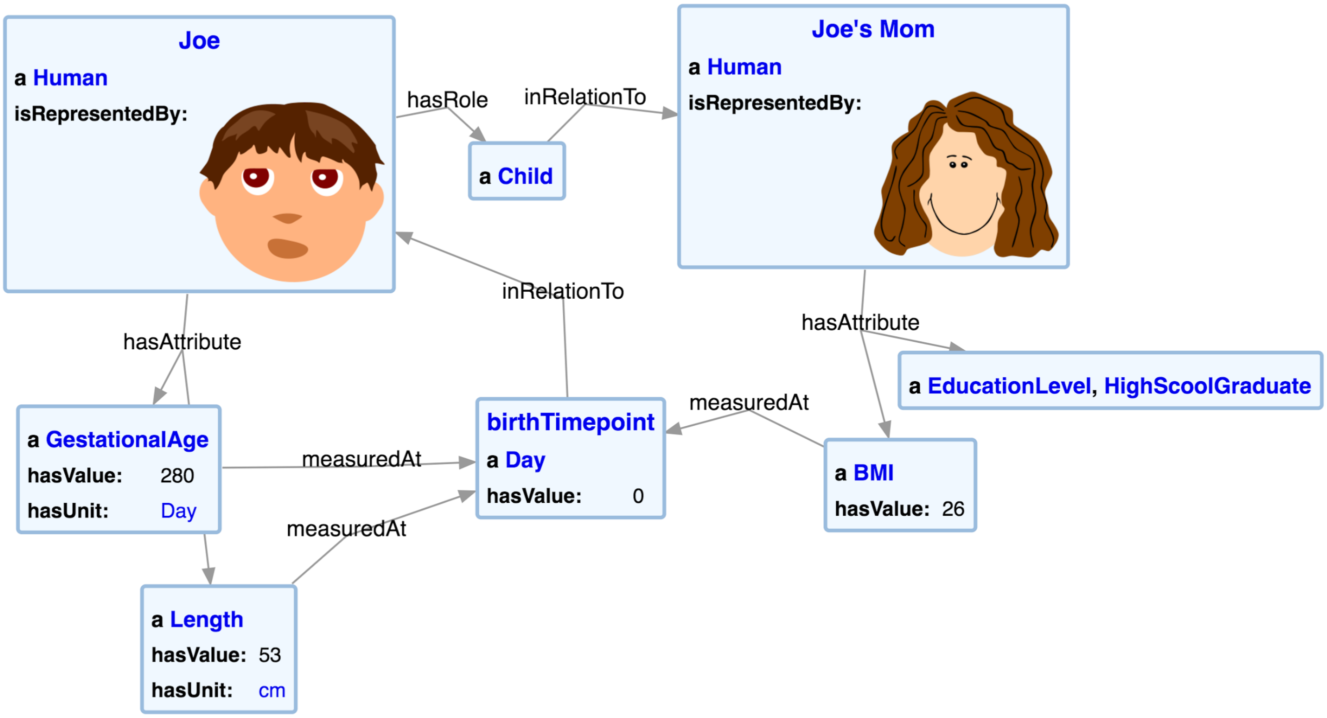

Figure 3: Time Interval Representation of Measurements. Most epidemiological attributes are measurements, here we show Gestational Age at birth, length at birth, and mother's BMI at birth. They are all related to the birth timepoint, measured in days. The mother's attribute of education level, being a High School graduate, is shown without a timepoint.

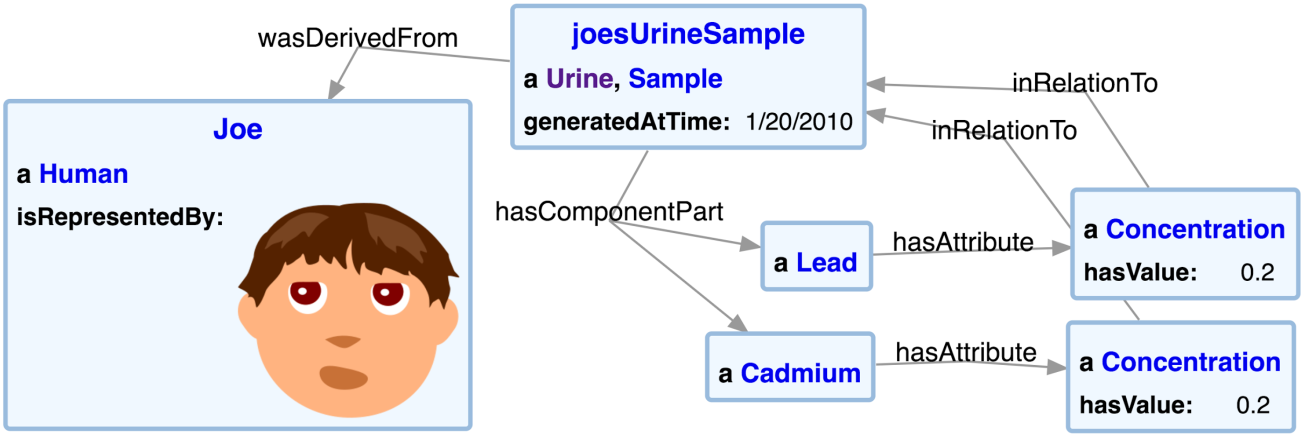

Figure 4: Representation of Analyte Concentrations in relation to a Human Derived Urine Sample. Most of the biomarker data in CHEAR is in the form of chemical analysis of elements and other analytes in blood, urine, and other sample types. We identify the component parts of the sample (Lead, Cadmium), and give them concentrations in relation to the sample itself, which was *prov:wasDerivedFrom* Joe.

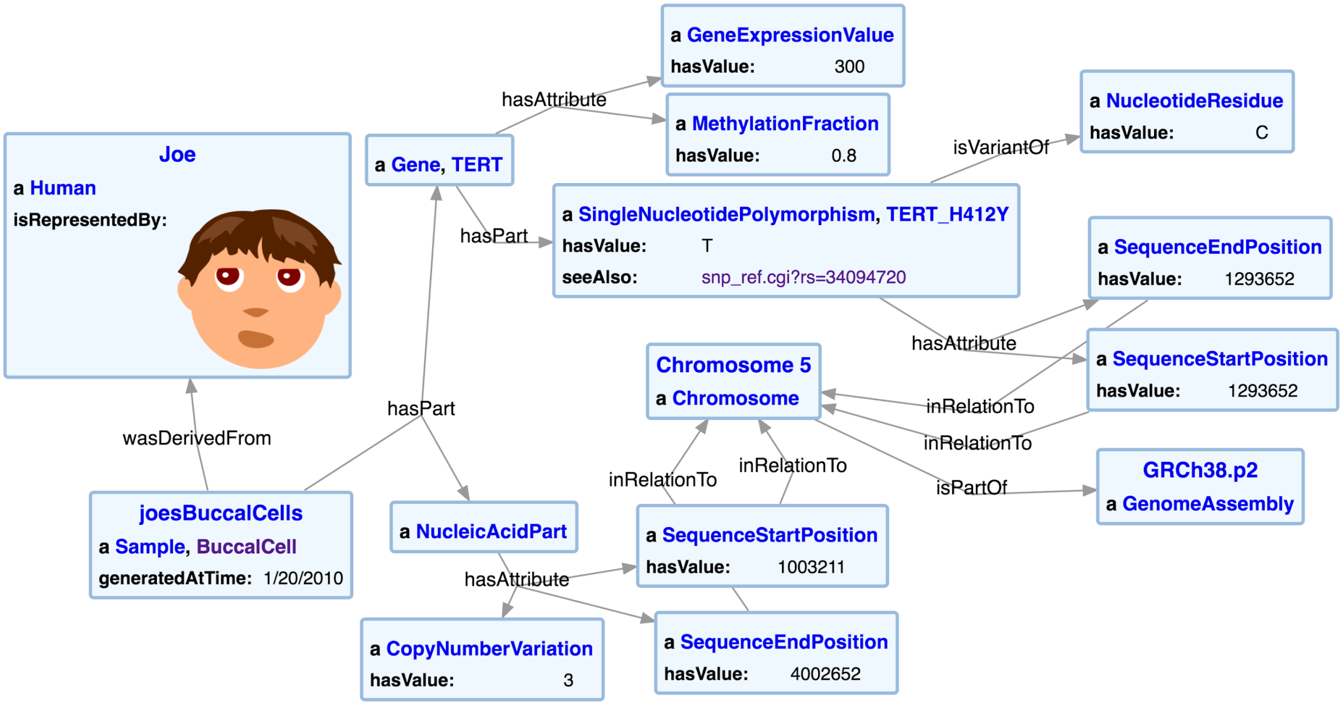

Figure 5: Representation of Genomic, Transcriptomic and Epigenomic Data. Finally, we are able to represent a wide array of genomic data using SIO's genomics modeling capabilities. Joe has a mutation on his TERT gene at on chromosome 5 at position 12993652, and that gene is part of a DNA region that has an extra copy on one chromosome. The data was against a Buccal Cell sample (often used for genomics analysis), derived from Joe.

We were also able to resolve differences in perspective between domains. For instance, in epidemiology, the term "metals" refers to any element that reacts biologically in the way that metals do. This therefore includes metalloids. We introduced a "metals an metalloids" class, into the CHEAR ontology, and since it is familiar to epidemiologists as "metals", we gave it the URI *chear:Metal*. It is not equivalent to *CHEBI:33521* (metal atom), but instead has metals and metalloids as subclasses. Additionally, we found it necessary to provide high-level classes for attributes, entities, and processes that relate to specific areas of interest within epidemiology. These classes are parallel to, but have no bearing on, the subsumption hierarchies of the domain-specific ontologies.

6 FUTURE WORK

We plan to build a much larger expansion of the CHEAR Ontology by attempting to model many key National Health and Nutrition Examination Survey (NHANES) [25] data dictionaries, which cover a number of epidemiological domains over many decades. NHANES includes datasets pertaining to other categories that are of interest to CHEAR research, including concepts related to dietary, laboratory and examination measures. We also plan to incorporate metabolite classes into the analyte hierarchy from the Reference Metabolite (RefMet) database from Metabolomics workbench.[10] Additional near term expansion areas for analytes include inflammation and oxidative stress.

One future goal is to create Semantic Data Dictionaries for the variables described in these datasets, in order to expand the range of concepts included in the CHEAR Ontology. A Semantic Data Dictionary will provide a formal means to map dataset columns into a compositional structure like the ones used in our examples. It will do so in a way that allows us to produce 1) OWL-based metadata for those datasets, creating explicitly defined classes that dataset columns map to, and 2) RDF data, or formal mappings of data to RDF, for the actual data that is described by the data dictionary that conforms to the OWL metadata we create. These Semantic Data Dictionaries are an expansion of the mapping process we used on the pilot data, and currently require human curation. For some studies, like NHANES, tools for web scraping can be used, such as the Python library Beautiful Soup [26], allowing for automatic population of variable names, labels, and definitions. Nevertheless, automating the population of entities, roles or relations that correspond to the variable cannot be accomplished simply by using web scraping techniques. Thus, an additional research direction is to leverage Natural Language Processing (NLP) methods to extract potential semantic qualities of variables from their label and description. Further, in preliminary discussions, we have found that it may be helpful to provide pre-populated starting points for the most commonly used data. Domain experts could then customize the mapping to their own studies. Further, we plan to produce a

[10] http://www.metabolomicsworkbench.org/data/refmet.php

number of Semantic Data Dictionaries to describe the data specifications that the CHEAR program is already developing for different kinds of laboratory analysis results, where appropriate.

7 CONCLUSIONS

We presented a new ontology for *in tela* integration of knowledge and data in global health and exposure research, called the CHEAR Ontology. We showed that it is possible and efficient to apply on-demand development approaches from agile software development (You Aren't Going to Need It) to developing broad, interdisciplinary ontologies. This process was facilitated through web-based collaboration of researchers from multiple domains. We were able to converge on a common data representation standard that covers the most commonly used data within the CHEAR program, with clear room for expansion to cover other types as they are needed. Using this approach, we generated an ontology that is ready for reuse that provides coverage in the areas of exposure and child health and that is compatible with many of the most widely used best practice ontologies and/or vocabularies in relevant areas. We are actively seeking partners and users for the ontology and welcome collaborators on the web-based semantic framework for on-demand ontology evolution and maintenance environments.

8 ACKNOWLEDGEMENTS

This work was funded by the National Institute of Environmental Health Sciences (NIEHS) Award 0255-0236-4609 / 1U2CES026555-01. We would like to thank Susan Teitelbaum at the Icahn School of Medicine at Mount Sinai for her leadership on the overall CHEAR data resource project, guidance in exposure and health domains, and feedback on the paper and ontology.

REFERENCES

[1] Consortium, W. W. W., et al. Json-ld 1.0: a json-based serialization for linked data.

[2] Courtot, M., Gibson, F., Lister, A. L., Malone, J., Schober, D., Brinkman, R. R., and Ruttenberg, A. Mireot: The minimum information to reference an external ontology term. *Applied Ontology 6*, 1 (2011), 23–33.

[3] Cristani, M., and Cuel, R. A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems (IJSWIS) 1*, 2 (2005), 49–69.

[4] Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research 36*, suppl 1 (2008), D344–D350.

[5] 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., and Hoehndorf, R. The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics 5*, 1 (2014), 14.

[6] Dumontier, M., and Hoehndorf, R. Realism for scientific ontologies. In *FOIS* (2010), pp. 387–399.

[7] Fenton, T., and Sauve, R. Using the lms method to calculate z-scores for the fenton preterm infant growth chart. *European journal of clinical nutrition 61*, 12 (2007), 1380–1385.

[8] Fox, P., McGuinness, D. L., Cinquini, L., West, P., Garcia, J., Benedict, J. L., and Middleton, D. Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience. *Computers & Geosciences 35*, 4 (2009), 724–738.

[9] Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. The units ontology: a tool for integrating units of measurement in science. *Database 2012* (2012), bas033.

[10] Horridge, M., and Patel-Schneider, P. F. Owl 2 web ontology language manchester syntax. *W3C Working Group Note* (2009).

[11] Hunt, A., and Thomas, D. The trip-packing dilemma [agile software development]. *IEEE Software 20*, 3 (May 2003), 106–107.

[12] Initiative, D. C. M., et al. DCMI metadata terms. *http://purl.org/dc/terms/* (2012).

[13] Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., McGarvey, P., and Gasteiger, E. Infrastructure for the life sciences: design and implementation of the uniprot website. *BMC bioinformatics 10*, 1 (2009), 136.

[14] JONES, D., BENCH-CAPON, T., and VISSER, P. Methodologies for ontology development. In *IFIP world computer congress* (1998), pp. 62–75.

[15] Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. Pubchem substance and compound databases. *Nucleic acids research* (2015), gkv951.

[16] Lebo, T., Sahoo, S., and McGuinness, D. PROV-O: The PROV Ontology. http://www.w3.org/TR/prov-o/, 2013.

[17] Lipscomb, C. E. Medical subject headings (mesh). *Bulletin of the Medical Library Association 88*, 3 (2000), 265.

[18] Mattingly, C. J., McKone, T. E., Callahan, M. A., Blake, J. A., and Hubal, E. A. C. Providing the missing link: the exposure science ontology exo, 2012.

[19] McCusker, J. P., Luciano, J. S., and McGuinness, D. L. Towards an ontology for conceptual modeling. In *ICBO* (2011).

[20] McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., et al. Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry 49*, 4 (2003), 624–633.

[21] McKinney, W., et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (2010), vol. 445, van der Voort S, Millman J, pp. 51–56.

[22] Miles, A., and Bechhofer, S. SKOS Simple Knowledge Organization System Reference, 2009.

[23] Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., and Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome biology 13*, 1 (2012), R5.

[24] Pinheiro, P., McGuinness, D. L., and Santos, H. Human-aware sensor network ontology: semantic support for empirical data collection. In *Proceedings of the 5th Workshop on Linked Science. Bethlehem, PA, USA* (2015).

[25] Pirkle, J. L., Brody, D. J., Gunter, E. W., Kramer, R. A., Paschal, D. C., Flegal, K. M., and Matte, T. D. The decline in blood lead levels in the united states: the national health and nutrition examination surveys (nhanes). *Jama 272*, 4 (1994), 284–291.

[26] Richardson, L. Beautiful soup documentation, 2007.

[27] Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research 40*, D1 (2012), D940.

[28] Snomed, C. Systematized nomenclature of medicine-clinical terms. *International Health Terminology Standards Development Organisation* (2011).

[29] Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research 37*, suppl 2 (2009), W623–W633.