Published on *Tetherless World Constellation* (https://tw.rpi.edu/web)

SemantEco Toxicity Extension

SemantEco Toxicity Extension

Contents

1. Intro and Purpose

The purpose of this extension is to allow users to learn about health effects of various contaminants measured in the SemantEco system. Based on symptoms, a user should be able to discover possible contaminants that are known to cause those symptoms. Similarly, by selecting a contaminant, a user can learn about the possible symptoms related to that exposure. This extension is still a work in progress, however here I will document my steps and thought process thus far, and my ideas for continuing work.

I have worked this semester to start an ontology to map Characteristics from the existing SemantEco system to the associated potential health effects. Right now, relationships in the ontology are implemented at a high level only, however I have attempted to structure my ontology so that more granularity can be added in the future. I have also worked on a use case for the related task of adding some search functionality of articles at Yale's Canary Database to our system; chemical name mappings from my ontology will facilitate this link. For the scope of the semester in Advanced Semantic Technologies, I focused on effects of metal poisoning, though other types of toxicity may be relevant to other projects as well: Canary Database examines effects of other contaminants, such as pesticides or infectious diseases, on species health. My classmate Brendan Ashby has been working with Darin Freshwater Institute, and measurements of interest to them include things like pH and dissolved oxygen, which are not necessarily toxic to humans.

My overall hope is that my ontology will provide the groundwork for an overall Health Facet that can provide both human health effect information from reputable sources such as the Center for Disease Control, and links to species health information on Canary Database as well. With this additional information, lay users may see how air and water quality measurements relate to them according to reliable institutions, while scientists and those with a deeper interest can also follow source links for additional information, including the full text of sources utilized.

2. Value of Semantics

The primary function of this extension is to provide more explicit information regarding what sorts of things Characteristics actually measure. The Characteristics as they exist in SemantEco right now are a flat list without any sort of hierarchy imposed to define them. The ChemicalCharacteristic subclass provides a place for all those Characteristics that are for chemicals - other subclasses could be added in the future for other types of Characteristics such as those for life form counts. This is ultimately to facilitate searching within all the Characteristics available in the system.

For scoping within the context of the Advanced Semantic Spring 2013 class, I have focused on toxicities related to a handful of metals. For each specific chemical, there are several different ways that measurements could be performed, as demonstrated by the multiple Characteristics that refer to one chemical. As far as the Toxicity Extension is concerned, a chemical in the water is a chemical in the water regardless of how the measurement is performed, and the health effects that result from exposure are not different based on how the measurement was done. (Health effects may differ as a result of the route of exposure, however this is specified separately). By providing a mapping of all Characteristics relevant to a chemical, to that chemical as a general category, the Toxicity Extension makes it easier to perform searches related to health effect information from data that only lists a chemical name. In turn, a hierarchy of chemicals (such as providing a Pesticide superclass for all pesticides, for example) would allow users to browse the Characteristics more easily, rather than having to slog through one flat list.

In order to facilitate further extension in the future, I made use of existing vocabulary in the ontology of Chemical Entity of Biological Interest (ChEBI). Further details of this implementation appear in the next section.

3. Ontology

SemantEco Toxicity Extension ontology available via github (RDF/OWL).

The data that exists in the SemantEco system currently includes numerous measurements from a variety of sources; one of the qualities of all of these measurements is the Characteristic. The Characteristic class name comes from the OBOE ontology, and refers to the quality being measured. For the moment, this is the only quality of the measurement that the Toxicity ontology is concerned with; measurements in SemantEco are of things in air or water, not in people themselves, and so we cannot make any assertions regarding exposure levels in any user. Not all Characteristics refer to things that are toxic or harmful – the full list includes things like 'Oxygen, dissolved percent saturation', 'Rainfall duration' and Color, in addition to things that are obvious health concerns like 'Arsenic, total dissolved (as As)'.

Although Protégé can alphabetize the list of all Characteristic classes, the list is otherwise completely unstructured. In addition, many Characteristics have names that look like they are meant to be identical (eg., 'Zinc, dissolved (as Zn)' versus 'Zinc, dissolved (as zn)', where the only difference is the capitalization of the element's symbol). All told, there are 9 different characteristics that represent measures of Zinc. While this may be an important distinction in other areas of SemantEco, for the purposes of health effects, a measure of Zinc contamination indicates that Zinc toxicity is a potential concern, regardless of how its presence in the water was confirmed.

For these reasons, I have started to implement some mappings of Characteristics to general chemicals using the Chemical Entity of Biological Interest (ChEBI) ontology. I created a class called ChemicalCharacteristic, under which I can file Characteristics that involve elements I am focusing on, to begin imposing some structure on the overall Characteristics list. ChemicalCharacteristic could be further refined in the future for better granularity; for example, it may be important to distinguish between organic and inorganic ChemicalCharacteristics.

ChemCharacteristic

> A ChemicalCharacteristic is both a Chemical and a Characteristic (subclass of both). This allows mapping from the measurement data in the SemantEco system to Chemicals as the Toxicity extension understands them.

Each ChemicalCharacteristic is also a subclass of a ChEBI:'chemical entity'; I selected the most general example for each element that still uniquely identifies that element for each ChemicalCharacteristic. For example, the 'zinc group molecular entity' contains subclasses for cadmium, mercury, and zinc molecular entities, so that class is not sufficient to describe any of them, including zinc itself. 'zinc molecular entity' contains subclasses for 'elemental zinc' as well as several zinc compounds, but these subclasses are not necessary to describe ZincCharacteristics. And so ZincCharacteristic is a subclass directly under 'zinc molecular entity'.

In the future, we may define an object property to specifically relate individual ChemicalCharactersitics to 'chemical entities' in a more granular way. For example, 'elemental arsenic' contains subclasses of specific arsenic ions and arsenic compounds. This may be important to distinguish in greater detail as the health effect information becomes more granular. ChEBI contains information relating roles of chemicals to biological processes, which may assist in describing health effects, since that ontology is ultimately concerned with how chemicals affect biology.

Toxicity-Cmap

> This diagram shows the relationships in the Toxicity extension, including how the ChemicalCharacteristic class links it to what exists already in SemantEco.

Chemicals are linked to Toxicities, which in turn have endpoint systems in the body, as defined by data on the Toxic Substances Portal maintained by the Agency for Toxic Substances & Disease Registry. I have also created classes for Health Effects and Organs, to achieve higher granularity in the future, to enable semantic reasoning to highlight which other organ Systems may be affected by contaminants by way of individual symptoms of Toxicity. Organs can be partonomically related to Systems; Systems can also be partOf other Systems.

One of the problems I encountered with constructing a representation of the human body systems in my ontology is the lack of universal agreement among medical professionals as to how systems are related. For example, the Disease Ontology lists hepatobiliary disease as a subclass of gastrointestinal system disease, whereas the National Library of Medicine defines the gastrointestinal system as exclusively including "the digestive structures stretching from the mouth to the anus", while not including accessory organs such as the liver or other biliary organs. I have chosen the second definition, with the GastrointestnalSystem and HepaticSystem separate, but both partOf the DigestiveSystem. Professor McGuinness also pointed out that there is a lot of debate over whether blood is considered an organ or a tissue or something else.

HealthEffects are also difficult because there is no universal standard for classification. It may or may not to be important to distinguish between symptoms (which tend to be subjective to the patient) and signs (which are objective, and detected during examination), but these distinctions are not always clear-cut. Additionally, disorders such as cancer can cause a variety of different signs or symptoms. I explored several different possible ways to structure the HealthEffects, including the ICD-10, but ultimately I chose to leave this part of the ontology alone for now. While the ICD-10 provides a potential hierarchy of symptoms, signs, and disorders, it is essentially meant to be a list of *codes* for diagnoses, rather than the diagnoses themselves.

I was also unable to find a good source of toxicity symptom data in machine-readable form. ATSDR has very thorough reports regarding health effects for many toxins, however these are all in PDF format. Even ATSDR's endpoint information is only written onto their pages, and so all the axioms I included in my ontology are ones I coded by hand.

Ultimately for future work, I would like to map HealthEffects to the Organs and organ Systems they afflict. This would allow a user to search for one or more HealthEffects, and the ontology would be able to determine possible sources of Toxicity based upon them. I would need to select maybe one authoritative source for one definite way to relate Organs to Systems, and Systems to other Systems. Based on Systems as Toxicity endpoints, the ontology could figure out possible sources of Toxicity, and their Chemicals. The reasoner would then be able to link the Chemicals to the SemantEco Characteristics by way of the ChemicalCharacteristic class, effectively joining the health effect data back to the measurement data.

Query-Structure
[example queries go here]
[results go here]

For example, a user wants to know all of the potential chemicals that can affect the nervous system. Selecting the nervous system in a dropdown facet menu on SemantEco would return a list of all potential Toxicities (from the handful of metal toxicities currently implemented, this would include Arsenic, Cadmium, and Lead Toxicities). From this result, the user could select one or more of these Toxicities in order to see all of the related measurements with Characteristics that fall into those ChemicalCharacteristic subclasses. To the user, this should appear seamless – it should be easy for a human user to infer how Lead Toxicity, Lead Characteristics, and Lead (Chemicals) are related, and so it is only necessary to specify how this logic flows in the ontology.

With further granularity, a user would also have the ability to select specific HealthEffects in order to query for potential Toxicities. Once the Toxicity Extension ontology is extended with axioms relating HealthEffects to Organs and Organs to Systems, then the reasoned could follow from a HealthEffect to the System it ultimately afflicts before continuing the logic mentioned above. The main difficulty in constructing the axioms necessary to show how HealthEffects afflict Organs or Systems is when the relationships are not obvious. Liver damage is clearly related to the liver, however something like abdominal pain, which is located near the gastrointestinal system but is essentially made up of impulses through the nervous system, is more difficult to relate to one single system. Also, something like a headache could have a variety of different causes depending on exactly what else is going on in the body. I would need to find authoritative sources for this information for all toxicities, ideally including ones that are comprehensive but largely in agreement on things such as the overall structure of the digestive system.

4. Canary Database

Also, by mapping Characteristics to ChemicalCharacteristics and 'chemical entities', the Toxicity extension provides an easy way to link SemantEco's data to Canary Database. Canary Database understands contaminants by their names (and refers to them as "Exposures"), not by their descriptions as Characteristics, and so this mapping would make it easier to search their articles based on a query containing a chemical name. They also have summaries for more general classes of contaminants, such as Pesticides. In this case, the Toxicity extension ontology could utilize superclass relationships to move up in the hierarchy to more general classes of Chemicals. This would also allow for easier browsing of ChemicalCharacteristics on the SemantEco side.

Given more time, I would also like to find or create an ontology to map species' common names to their scientific names. Right now, SemantEco only makes use of scientific names for birds via the eBird/clements ontology; providing mappings to common names would make the system easier for the lay user to search. Additionally, Canary Database currently has the ability to search articles by species, however some species are listed by their scientific name, and some by their common name; however only one or the other seems to appear in the "browse by species" list. For example, mallard ducks appear in that list only by their common name, and do not also appear in the species list as *Anas platyrhynchos*. The ability to map between both names would facilitate search on both sites.

In addition, it would be useful to include other species. Brendan's data from Darin Freshwater Institute includes data on aquatic species, such as plankton and crustaceans. While many of these do not have common names, utilizing a taxonomy ontology of some sort would be able to make the implicit taxonomic information explicit to assist in filtering for search, both within SemantEco and when linking out to Canary Database.

Brendan and I have both looked at NCBI organismal classification (NCBI Taxon) for species taxonomy information, and, while this currently remains unimplemented, this ontology appears to be suitable for our needs.

5. Provenance

Provenance remains very important throughout this extension, as it is necessary that the health information is trustworthy and authoritative. So far, my primary source has been the ATSDR, part of the Center of Disease Control. I have also looked into some toxicity information via the National Institutes of Health, but I have not encoded any of that data in my ontology thus far.

There are disagreements in some medical definitions, as mentioned previously. However I do not anticipate these distinctions will be important to the lay user, nor will utilizing, for example, a different definition of "digestive system" render the system unusable to experts who may define it differently. It is only important that the definition choices are consistent throughout the extension, and the source(s) for those definitions are documented and authoritative. Filtering searches based on differing sources for the definition of "blood", for example, will only make the extension more confusing.

ATSDR publishes Minimal Risk Levels for each contaminant, defined as "an estimate of the daily human exposure to a hazardous substance that is likely to be without appreciable risk of adverse non-cancer health effects over a specified duration of exposure". If other authoritative health sources publish different levels that are later incorporated into this extension, then it may make sense to offer filtering based on these. This would parallel differing water quality definitions between state and federal governments as already implemented in SemantEco. However, the data measurements in SemantEco are in effluence; a limitation of the information presented in this Toxicity extension is that it assumes exposure has occurred based on those measurements. While it may be useful to present these risk levels to the user, it would not make sense to filter searches in this manner.

A possible future use of provenance filtering could be the ability to filter health effects based on age, for cases where these effects are different in children than in adults.

6. Overview of Limitations and Next Steps

Because the Minimal Risk Levels (MRLs) are not implemented in the health ontology, the Toxicity Extension is not able to reason whether or not an exposure event has definitely occurred. Data from the United States Geological Survey (USGS) is for measurements in the water or air, but we have no reliable way to relate this to whether or not a user personally has experienced hazardous levels of exposure.

I expect we will never have this capability necessarily, as the health extension is not meant to replace a doctor's diagnosis. The overall health facet should ultimately be able to provide information regarding known symptoms of various toxicities. But we can still provide information detailing what levels are hazardous according to the ATSDR, along with the other ATSDR information regarding differential endpoints based on route and duration of exposure to the contaminant. I implemented classes for ExposureRoute and ExposureDuration so that this can be further applied for more granularity regarding specific toxicities in the future.

The ChEBI ontology provides a very granular model for relating 'chemical entities' to their parts and to the roles they have biologically. As the SemantEco Toxicity extension aggregates more data regarding how toxic effects are induced, the ontology could further be leveraged to detail the effects in the human body in the presence of these contaminants. This could be useful for building symptoms upwards from a cellular level: for example, providing details as to what processes lead to a non-specific symptom such as a headache for the various toxicities that induce that symptom. With biological roles and effects defined at a very low level, it may also be possible to apply them to other species, as well.

Advanced SemTech 2013

Source URL: https://tw.rpi.edu/web/SemantEco-Toxicity-Extension