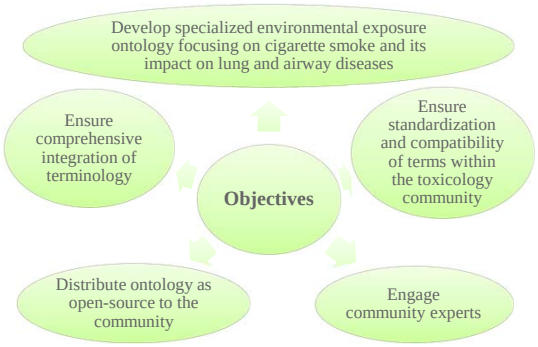
CSEO – The Cigarette Smoke Exposure Ontology

Sam Ansari[1], Erfan Younesi[2], Michaela Guendel[2], Shiva Ahmadi[2], Chris Coggins[3], Julia Hoeng[1], Martin Hofmann-Apitius[2], Manuel Peitsch[1]
[1] Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
[2] Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[3] Carson Watts Consulting, 1266 Carson Watts Rd, King, NC 27021-7453, US

Abstract

In the past years, significant progress has been made in the development and use of experimental settings for collection of experimental data on tobacco exposure and the diseases induced by it. Despite the growing number of such data, there has been no community wide effort to facilitate the centralization and integration of tobacco exposure data scattered throughout a range of disparate sources. Moreover, to fulfill the aim of exposure and disease impact studies, it is of utmost importance to more reliably and efficiently establish the causal link to disease. Ontologies are structural frameworks for organizing knowledge, enabling information retrieval, and supporting data integration, data analysis, and exchange of knowledge within the community. Cigarette Smoke Exposure Ontology (CSEO) was developed which is a specialized ontology with particular focus on the cigarette smoke exposure and related various experimental systems. Combining efforts of domain experts as well as novel computational methods, the ontology successfully describes exposure related terminology ranging from the scope and design definition of an experiment to its outcome with link to molecular events and ultimately to disease. After several iterations between the computational and domain expert groups, CSEO encompasses more than 20.000 concepts and classes. This ontology is represented in web ontology language (OWL) format and is made freely available to the community through several channels.

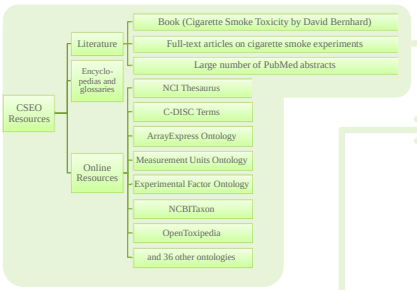
Objective

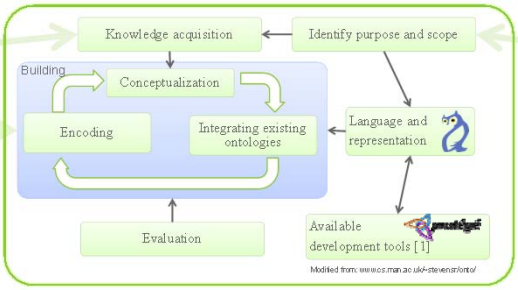


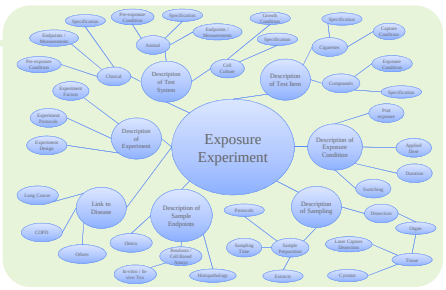
Purpose

- Formalization of the community view: agreement on a single controlled vocabulary
- Enabling Interoperability: supporting communication between different systems
- Model-based knowledge acquisition: to acquire knowledge in a structured way
- Re-use of existing domain knowledge: to avoid re-inventing the wheel
- Ontology-driven text-mining: semantically enhanced information retrieval and extraction

Procedure

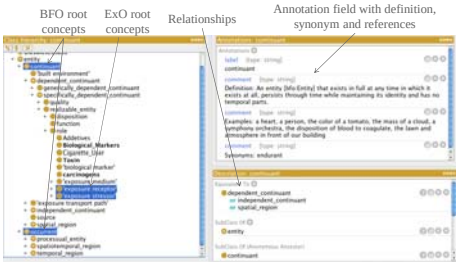


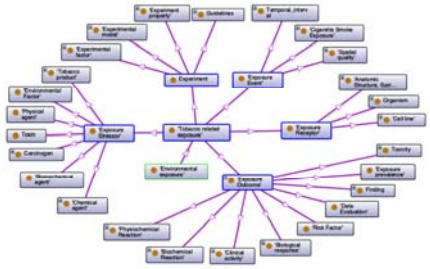




Combining efforts of domain experts as well as novel computational methods, the ontology describes exposure related terminology ranging from the scope and design definition of an experiment to its outcome with link to molecular events and ultimately to disease. During the process of ontology building, a hybrid approach combining both bottom-up and top-down methods was used, so that the ontology was populated at the level of superclasses and subclasses simultaneously. Development of CSEO was accomplished in four phases according to the standard life cycle of the ontology building [5].

Results





Second-level concept structure of CSEO

Structural evaluation

As structural statistics indicate, the high number of classes and leaves together with high values for average width and the fanout factor point towards a broad coverage of concepts by the ontology, whereas the values for depth show specificity of the concept types to the domain of cigarette smoke exposure risk.

number of classes	20028
number of leaves	14356
max width/breadth	2504
avg. width/breadth	953.72
max depth	21
total no. of children	26301
avg. number of children	1.32
avg. depth (avg. root-to-leaf distance)	11.32
depth variance (var(d) = E[d^2]-E[d]^2)	7.76
width/breadth variance (var(w) = E[w^2]-E[w]^2)	894681.06
tangledness (no. nodes with 2+ parents / total no. nodes)	0.27
fanout factor (no. leaf classes / number classes)	0.71

Functional evaluation

The result of this evaluation shows that the ontology in its current form is able to capture a wide range of cigarette smoke exposure concepts in the knowledge domain of exposure with a reasonable sensitivity and specificity towards manual curation.

	Precision	Recall	F-score
CSEO performance	69.23%	77.81%	73.26%

Usability and access

Since only a proper documentation can ensure direct access and efficient usability of the ontology, we have created a wiki page that contains instructions for using the ontology, documentation on purpose and scope of the ontology, as well as information on interfacing the ontology. The wiki is accessible through the following link:

https://publicwiki-01.fraunhofer.de/CSEO-Wiki/index.php/Main_Page

CSEO will further be made available through NCBO Bioportal: **http://bioportal.bioontology.org/**

Expert evaluation

Experts in the knowledge domain of cigarette smoke exposure risk were asked to design several complex questions to be posed to the ontology. As an example, the following question was considered to test the performance of the ontology:

What are the potential effects of the toxicity induced by tobacco smoke constituents on smokers?

Query: (([CSEO:"Smoke Constituent"]) AND [CSEO:"Toxicity"]) AND [CSEO:"Tobacco"]

Total number of retrieved documents: 21
Date: 21.03.2013
Number of relevant documents containing correct answer: 17
Percentage of correct retrieval: 81%

PMIDs of relevant documents manually curated for correct answer: 14521141 , 1188959 , 18848577 , 21651432 , 17661226 , 2002748 , 12857635 , 19330121 , 14698566 , 11731039 , 18383128 , 16859820 , 21651433 , 21417965 , 21651431 , 15072838 , 18464053

Conclusion

CSEO aims to provide a standard platform for interoperability of the environmental exposure knowledge and has potential to become a widely used standard within the academic and industrial community. Mainly because of the emerging need of the 21st century of toxicology to controlled vocabularies as well as the lack of suitable ontologies for this domain, CSEO is preparing the ground for a sustainable ontology within the exposure domain and will also enable the scientific community to compare their findings and therefore increases scientific transparency. The availability of this ontology will further improve the specificity of ongoing knowledge engineering activities by improving text mining and knowledge representation. Ontology-driven text mining is a relatively emerging topic, which takes advantage of extensive dictionary of knowledge domain concepts to mine the textual body of the literature in a systematic manner.
This ontology contains a total number of 20086 concept classes and is represented in web ontology language (OWL) format and is made freely available to the community through several channels.

Reference

[1] T. Tudorache, C.I. Nyulas, N.F. Noy, M.A. Musen. WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web, Semantic Web Journal, IOS Press, 2011.
[2] Smith B, Kumar A, Bittner T. Basic formal ontology for bioinformatics. Journal of Information Systems, 2005.
[3] Mattingly CJ, McKone TE, Callahan MA, et al. Providing the missing link: the exposure science ontology ExO. Environ. Sci. Technol. 2012;46:3046–3053.
[4] Smith B, Ceusters W, Klagges B, Kohler J, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C. Relations in Biomedical Ontologies. Genome Biology, 2005, 6:R46.
[5] Stevens, R., C.A. Goble, and S. Bechhofer, Ontology-based knowledge representation for bioinformatics. Brief Bioinform, 2000. 1(4): p. 398-414.

PMI RESEARCH & DEVELOPMENT
The 6th International Biocuration Conference
Churchill College, Cambridge, UK
April 7-10, 2013

Philip Morris International Research & Development, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
T: +41 58 242 21 11, **F**: +41 58 242 28 11, **W**: www.pmi.com