Proceedings of the 13th ISMB'2010 SIG meeting "Bio-ontologies" 2010:120-123.

The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench

Michel Dumontier [a], Bosse Andersson [b], Colin Batchelor [c], Christine Denney [d], Christopher Domarew [e], Anja Jentzsch [f], Joanne Luciano [g], Elgar Pichler [h], Eric Prud'hommeaux [i], Patricia L. Whetzel [j], Olivier Bodenreider [k], Tim Clark [l], Lee Harland [m], Vipul Kashyap [n], Peter Kos [l], Julia Kozlovsky [p], James McGurk [o], Chimezie Ogbuji [q], Matthias Samwald [r], Lynn Schriml [s], Peter J. Tonellato [l], Jun Zhao [t], Susie Stephens [u]*

[a]Carleton University, Ottawa, Canada. [b]AstraZeneca, Lund, Sweden. [c]Royal Society of Chemistry, Cambridge, UK. [d]Eli Lilly, Indianapolis, IN, USA. [e]Warrington Hospital, Warrington, UK. [f]Freie Universitat, Berlin, Germany. [g]Predictive Medicine Inc., Belmont, MA, USA. [h]W3C HCLSIG. [i]W3C, Cambridge, MA, USA. [j]Stanford University, Stanford, CA, USA. [k]National Library of Medicine, Bethesda, MD, USA. [l]Harvard Medical School, Cambridge, MA, USA. [m]Pfizer, Sandwich, UK. [n]Cigna, Hartford, CT, USA. [o]Daiichi Sankyo, NJ, USA. [p]AstraZeneca, Waltham, MA, USA. [q]Cleveland Clinic, Cleveland, OH, USA. [r]Digital Enterprise Research Institute, Galway, Ireland. [s]University of Maryland, Institute for Genome Sciences. [t]University of Oxford, Oxford, UK. [u]Johnson & Johnson Pharmaceutical Research & Development L.L.C., Radnor, PA, USA.

ABSTRACT

The Translational Medicine Ontology provides terminology that bridges diverse areas of translational medicine including hypothesis management, discovery research, drug development and formulation, clinical research, and clinical practice. Designed primarily from use cases, the ontology consists of essential terms that are mapped to other ontologies. It serves as a global schema for data integration while simultaneously facilitating the formulation of complex queries across heterogeneous sources. We demonstrate the utility of the ontology through question answering over a prototype knowledge base composed of sample patient data integrated with linked open data. This work forms a basis for the development of a computational platform for managing information relevant to personalized medicine.

1 INTRODUCTION

Personalized medicine aims to identify effective therapeutic regimes that are safe and effective (Trusheim et al. 2007). Essential to the realization of personalized medicine is the development of information systems capable of providing accurate and timely information about patients, drugs and therapeutic treatments. Integration of a patient's electronic health record (EHR) with publicly accessible information creates new opportunities for clinical research and patient care. EHRs encourage the identification of adverse events and outbreak awareness and serve as a rich set of longitudinal data, from which researchers can study disease, co-morbidity, and treatment outcome. While supplying patient data to the scientific community presents both technical and social challenges (Rodwin 2009), a comprehensive system that maintains individual privacy but provides a platform for the analysis of the full extent of patient data is vital for personalized treatment and objective prediction of drug response (Roses 2008). The impetus to collect and disseminate relevant patient specific data for use by clinicians, researchers, and drug developers has never been stronger.

Inability to access medical records is only partly responsible for the suboptimal use of data. Large quantities of fragmented and unstructured information prohibit physicians and researchers from easily gaining insight from clinical encounters, or obtaining the up-to-date evidence-based guidelines for disease diagnosis and treatment. Such complexity can impair the clinician's ability to accurately and rapidly prescribe drugs that are safe and effective for the patient, and covered by the patient's insurance provider. Translational medicine depends on the comprehensive integration of the entire breadth of patient data to facilitate and evaluate drug development (Woolf 2008). Ontologies are expected to play a major role in the automated integration of patient data with relevant information to facilitate discovery research, hypothesis management, formulation, clinical trials, and clinical research.

Semantic Web technologies enable the integration of heterogeneous data using explicit semantics, the expression of rich and well-defined models for data aggregation, and the application of logic to gain new knowledge over the raw data. The four main Semantic Web standards for knowledge representation are: Resource Description Framework (RDF); RDF Schema (RDFS); Web Ontology Language (OWL); and SPARQL as a query language. OWL ontologies have been developed to support drug, pharmacogenomic and clinical trials (Dumontier & Villanueva-Rosales 2009)(Coulet et al. 2006)(Arikuma et al. 2008) are increasingly used in the health care and life sciences (Shah et al. 2009).

In this paper, participants in the Translational Medicine Ontology task force of the World Wide Web Consortium's Health Care and Life Sciences Interest Group, present the Translational Medicine Ontology (TMO). Developed over

* To whom correspondence should be addressed.
michel_dumontier@carleton.ca;

the course of a year largely through weekly teleconference calls, the TMO bridges existing open domain ontologies and provides a framework to relate and integrate patient-centric data from the entire bench-bedside translational enterprise.

2 USE CASE

Alzheimer's Disease (AD) is an incurable, degenerative, and terminal disease with few therapeutic options. AD is influenced by a range of genetic, environmental and other factors. Identification of prognostic biomarkers would significantly impact and guide the diagnosis, prescription, and development of therapeutic agents would significantly impact future practice. Efficient aggregation of relevant information to help understand the pathology would benefit researchers, clinicians, and patients and would also facilitate the development of target compounds to reduce or even prevent the burden of the disease. We demonstrate the value of TMO by aggregating relevant semantically annotated AD data of interest from multiple data sources.

3 METHODS

3.1 Ontology Design

The TMO was built with Protégé 4.0.2 and represented as an OWL2 compliant ontology. Terms are defined in the http://www.w3.org/2001/sw/hcls/ns/transmed/ namespace. The ontology is available from the Google Code project http://code.google.com/p/translationalmedicineontology/.

TMO terms were initially obtained from a lexical analysis of questions that might be posed to 16 types of users involved in research, clinical care and business (**Table 1**).

Table 1 Users and their interests in translational medicine

Category	User	Interest
Research	Biologist (*in vivo*, *in vitro*, cellular & molecular)	Target identification, assay development, target validation
	Bioinformatician	Biological knowledge management, cellular modeling
	Immunologist	Natural defense mechanisms
	Cheminformatician	Predictive chemistry
	Medicinal chemist	Drug efficacy
	Systems physiologist	Tolerance, adverse events
Clinic	Clinical trial specialist	Trial formulation, recruitment
	Clinical decision support	Data analysis, trend finding
	Primary care physician	General, conventional care
	Specialty medical provider	Specialized treatments
Business	Sales & marketing	Revenue generation
	Strategic/portfolio manager	Assessing market opportunities
	Project manager	Prioritizing resources & activities
	Health plan provider	Insurance coverage

source: http://esw.w3.org/topic/HCLSIG/PharmaOntology/Roles

The TMO defines 75 classes spanning material entities (e.g. molecule, protein, cell lines, pharmaceutical preparations), roles (e.g. subject, target, active ingredient), processes (e.g. diagnosis, study, intervention), and informational entities (e.g. dosage, mechanism of action, sign/symptom (Scheuermann et al. 2009), family history). The TMO extends the basic types defined in the Basic Formal Ontology and uses relations from the Relation Ontology.

In TMO, entities that appear in statements that hold in general (e.g. 'patients participate in consultations' and 'active ingredient is a role played by a molecular entity') form key background knowledge and are captured as classes in the ontology. In contrast, particulars (e.g. 'a patient with a given name' and 'a blister package of a pharmaceutical product with a particular identifying code on it') instantiate classes in the ontology. These rules apply to material entities as well as roles and processes. Consequently, a particular consultation at a given time and day; the particular patient role in that consultation; and the physician role in that consultation are each instantiations of a class.

Table 2 Representative mappings between TMO and target terms

Label	TMO	Target
Protein	0035	ACGT:Protein, BIRNLex:23, CHEBI:36080, FMA:Protein, GO:0003675, GRO:Protein, Galen:Protein, NCIt:Protein, PRO:000000001, SNOMEDCT:88878007, SO:0000358, UMLS:C0033684
Gene	0037	FMA:Structural_gene, GRO:Gene, Galen:Gene, LNC:LP32747-5, MSH:D005796, NCIt:Gene, NCIt:Gene_Object, NDFRT:C242394, PRO:Gene, SNOMEDCT:67271001, SO:0000704, UMLS:C0017337
Diagnosis	0031	ACGT:Diagnosis, FHHO:Diagnosis, Galen:Diagnosis, LNC:LP72437-4, MSH:D003933, NCIt:Diagnosis, OBI:0000075, OCRe_clinical:Diagnosis, SNOMEDCT:439401001, UMLS:C0011900
Disease	0047	ACGT:Disease, BIRNLex:11013, DOID:4, GRO:Disease, LNC:LP21006-9, MSH:D004194, NCIt:Disease_or_Disorder, NDFRT:C2140, OBI:0000155, UMLS:C0012634

Abbreviations: ACGT- ACGT Master Ontology, BIRNLex – BIRNLex Ontology, CHEBI – Chemical Entities of Biological Interest, CTO – Clinical Trial Ontology, DOID – Human Disease Ontology, FMA – Foundation Model of Anatomy, FHHO – Family Health History Ontology, Galen – Galen Ontology, GO – Gene Ontology, GRO – Gene Regulation Ontology, LNC – Logical Observation Identifer Names and Codes, MSH- Medical Subject Headings, NCIt – NCI theraurus, NDFRT – National Drug File, OBI – Ontology for Biomedical Investigation, OCRe_- Ontology for Clinical Research, PATO – Phenotypic Quality Ontology, PRO – Protein Ontology, SNOMED-CT, SNOMED clinical terms, SO – Sequence Ontology, UMLS – Unified Modeling Language System.

Additional terms were obtained by a cursory analysis of the types referred to by the linked open data (Section 3.2). Polysemous terms were disambiguated into separate entities. For instance, a "drug" can refer to the whole pharmaceutical product, or just the active ingredient. The TMO differentiates these meanings as a "molecular entity" (TMO_0034)

for single molecule kinds, "active ingredient" (TMO_0000) for biologically active chemicals in formulated pharmaceuticals, "formulated pharmaceutical" (TMO_0001) for a substance that may or may not have been approved by a regulatory authority, and "pharmaceutical product" (TMO_0002) for a drug approved by a regulatory authority.

To establish the TMO as a global ontology, we created 223 class equivalence mappings (using *owl:equivalentClass*) from 60 TMO classes to 201 target classes from 40 ontologies (see **Table 2**). Initially identified using the NCBO Bioportal and UMLS, each mapping was manually validated.

3.2 Data Sources

The data sources used in this study include formulary lists, pharmacogenomics information, clinical trial lists, and scientific data about marketed drugs (**Table 3**).

Table 3 Data sources used in this study

LODD	Prefix	Dataset	Description
x	linkedct	Clinicaltrials.gov	Registry of clinical trials
	dubois	AD diagnostic	AD diagnostic criteria
x	dailymed	DailyMed	Marketed & FDA approved drugs
x	diseasome	Diseasome	The genetic basis of disease
x	drugbank	DrugBank	Detailed drug data & drug target
x	medicare	Medicare	Medicare D approved drugs
	pchr	Patient	Fake patient data
	pharmgkb	PharmGKB	Drug response to genetic variation
x	sider	SIDER	Side effects of marketed drugs

LODD – 'x' indicates a linking open drug data dataset

All datasets except for PharmGKB, diagnostic criteria and patient records are available through the Linking Open Drug Data (LODD)[1] project (Jentzsch et al. 2009). Alzheimer's diagnostic criteria were obtained from Dubois *et al.* (Dubois et al. 2007). Seven patient health records were manually created to capture demographic information, contact information, family history, life style data, allergies, immunizations, information on conditions, procedures, prescriptions, and encounters with members of the medical community. The patient record was defined by an XML schema, based in part on the Indivo schema[2], and converted into RDF using an XSL stylesheet.

3.3 Data Mapping

Mappings between LODD datasets were generated with LinQuer (Hassanzadeh et al. 2009) for resources with non-identical labels, and Silk (Volz et al. 2009) which employs similarity metrics including string, numeric, data, URI, and set comparison methods. Entity identity was asserted using *owl:sameAs*. The mappings were augmented by those provided for PharmGKB via Bio2RDF (Belleau et al. 2008). 25 Mappings between LODD dataset types and TMO types were established using *owl:equivalentClass*.

[1] http://esw.w3.org/HCLSIG/LODD/Data

[2] http://wiki.indivohealth.org/index.php/Main_Page

3.4 Translational Medicine Knowledge Base

The Translational Medicine Knowledge Base (TMKB) is an RDFS-reasoning capable Semantic Web knowledge base composed of the TMO, RDFized datasets, and equivalence mappings. Files were loaded into OpenLink Virtuoso 6 open source community edition, which provides a SPARQL endpoint[3] and a faceted text search interface[4].

Table 4 Questions and answers using TMO-integrated data sources

Question	Answer
Clinic	
What are the diagnostic criteria for AD?	There are 12 diagnostic inclusion criteria and 9 exclusion criteria.
Does Medicare D cover Donepezil?	Medicare D covers 2 brand name formulations of Donepezil: Aricept and Aricept ODT.
Have any AD patients been treated for other neurological conditions	Patient 2 was found to suffer from AD and depression.
Clinical Trial	
Since my patient is suffering from drug-induced side effects for AD treatment, identify an AD clinical trial with a different mechanism of action (MOA)	Of the 438 drugs linked to AD trials, only 58 are in active trials and only 2 (Doxorubicin and IL-2) have a documented MOA. 78 AD-associated drugs have an established MOA.
Find AD patients without the APOE4 allele as these would be good candidates for the clinical trial involving Bapineuzumab?	Of the four patients with AD, only one does not carry the APOE4 allele, and may be a good candidate for the clinical trial.
What active trials are ongoing that would be a good fit for Patient 2?	58 Alzheimer trials: 2 mild cognitive impairment, 1 hypercholesterolaemia, 66 myocardial infarction, 46 anxiety, and 126 depression.
Research	
What genes are associated with or implicated in AD?	Diseasome and PharmGKB indicate at least 97 genes have some association with AD.
Which SNPs may be potential AD biomarkers?	PharmGKB reveals 63 SNPs
Which market drugs might potentially be re-purposed for AD because they modulate AD implicated genes?	57 compounds or classes of compounds that are used to treat 45 diseases, including AD, hyper/hypotension, diabetes and obesity.

4 RESULTS

Translational medicine is facilitated when the full extent of patient data is integrated and bench-to-bedside data-dependent questions can be asked and answered. Here, we focus on questions that a physician within clinical practice would like to have answered more easily, such as the diagnosis of a disease and the prescription of drugs. However, TMO has the potential to be equally relevant to scientists developing new pharmaceutical products. While simple questions may be answered by queries on a single data set, other scientific questions may only be addressed when di-

[3] http://tm.semanticscience.org/sparql

[4] http://tm.semanticscience.org/fct

verse data sets are fully integrated. Importantly, answering more sophisticated questions may require inference over i) the subclass hierarchy of TMO types or ii) through equivalence mappings. Examples of queries that can now be executed with SPARQL[5] are listed in **Table 4**.

5 CONCLUSION

The TMO aims to support translational medicine by providing terms that facilitate interoperability for information stemming from the bedside to the bench. Our AD-focused use case demonstrates the use of TMO in translational research in the context of a well known disease. While the medical history of our patients is not extensive, it reflects the reality of incomplete medical records. Consistently implemented EHRs will play an ever more important role in broader professional collaborations between researchers and clinicians. More effective integration of data, as we have demonstrated here through the use of formal ontologies, should enable pattern recognition in a clinical setting to identify superior efficacy of certain drugs over others in specific sections of the population. For example, a clinician would be able to obtain a list of effective, safe, evidence-base therapies for administration to a specific patient and was also available under the patient's health plan.

Since our work specifically focused on integrating existing datasets using a common vocabulary, we invariably acquired terms that are either ontologically difficult or were not found in existing community ontologies. The term "side effect" is particularly challenging because they are so varied. For instance nightmares are processes, but tender gums are dispositions that are realized in processes (sensation of pain in gums when palpated), or the qualities of material entities (coldness of extremities). While the TMO has 'adverse drug event' (TMO_0043), it will take additional effort to correctly assign SIDER-listed side effects.

Future TMO work will focus on the addition of entities related to drug discovery and drug development in order to increase its utility for the pharmaceutical industry. Another key goal is the development of a role-based user interface that would encourage vendors of EHRs to use ontologies such as the TMO to not only guide question answering, but also improve representation and integration of data. TMO is a first step towards enabling scientists to use systems to reason across vast amounts of health care and life science data.

ACKNOWLEDGEMENTS

We would like to thank the National Center for Biomedical Ontology (NCBO), the entire Translational Medicine Ontology Task Force, and the support provided by the World Wide Web Consortium (W3C) to the W3C Health Care and Life Science (HCLS) Interest Group.

[5] http://esw.w3.org/topic/HCLSIG/PharmaOntology/Queries

REFERENCES

Arikuma, T, S Yoshikawa, R Azuma, K Watanabe, K Matsumura, and A Konagaya. 2008. Drug interaction prediction using ontology-driven hypothetical assertion framework for pathway generation followed by numerical simulation. *BMC Bioinformatics* 9 Suppl 6: S11.

Belleau, F, M-A Nolin, N Tourigny, P Rigault, and J Morissette. 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Semantics.* 41,5: 706-716.

Coulet, A, MM Smail-Tabbone, A Napoli, and M-D. Devignes. 2006. Suggested Ontology For Pharmacogenomics (SO-Pharm): Modular Construction And Preliminary Testing. *Lecture Notes in Computer Science.* 4277: 648-657.

Dubois, B, HH Feldman, C Jacova, ST DeKosky, P Barberger-Gateau, J Cummings, A Delacourte, D Galasko, S Gauthier, and G Jicha. 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology* 6, no. 8: 734-746.

Dumontier, M, and N Villanueva-Rosales. 2009. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics* 10, no. 2: 153-163.

Hassanzadeh, K, A Kementsietsidis, L Lim, R J Miller, and M Wang. 2009. A framework for semantic link discovery over relational data. In *CIKM '09: Proceeding of the 18th ACM Conference on Information and Knowledge Management*, 1027-1036. New York, NY, USA: ACM.

Jentzsch, A, J Zhao, O Hassanzadeh, K-H Cheung, M Samwald, and B Andersson. 2009. Linking open drug data. In *Triplification Challenge of the International Conference on Semantic Systems*. Graz, Austria.

Rodwin, MA. 2009. The case for public ownership of patient data. *JAMA: The Journal of the American Medical Association* 302, no. 1: 86-8.

Roses, A D. 2008. Pharmacogenetics in drug discovery and development: a translational perspective. *Nature Reviews. Drug Discovery* 7, no. 10: 807-17.

Scheuermann, RH, W Ceusters, and B Smith. Toward an Ontological Treatment of Disease and Diagnosis. In *Proceedings of the Second AMIA Summit on Translational Bioinformatics*. San Francisco, CA.

Shah, N H, C Jonquet, A P Chiang, A J Butte, R Chen, and M A Musen. 2009. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics* 10 Suppl 2: S1.

Trusheim, M R, E R Berndt, and F L Douglas. 2007. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature reviews. Drug discovery* 6, no. 4: 287-93.

Volz, J, C Bizer, M Gaedke, and G Kobilarov. 2009. Silk–a link discovery framework for the web of data. *2nd Linked Data on the Web Workshop LDOW2009* 42, no. 4: 817-819.

Woolf, S H. 2008. The meaning of translational research and why it matters. *JAMA: The Journal of the American Medical Association* 299, no. 2: 211-3.