ONTOLOGY BASED DATA AND INFORMATION INTEGRATION IN BIOMEDICAL DOMAIN

Assist. Prof. Dr. Gocheva D. G., Assist. Eng. Eminova H. M., Prof. Dr. Batchkova I. A.
Dept. of Industrial Automation, University of Chemical Technology and Metallurgy
Bul. Kl. Ohridski 8, Sofia, Bulgaria

dani@uctm.edu, h_eminova@uctm.edu, idilia@uctm.edu

Abstract: *One of the main problems of biomedical informatics in the effort to increase its contribution in knowledge retrieval and decision making is the integration of ever-increasing amounts of information and data from multiple heterogeneous sources and domains - clinical, medical, biological etc. The paper proposes an ontology based approach for integration of biomedical data and information using the Linked Open Data vocabularies and a D2RQ-mapped database. A simple example of semantic integration of heterogeneous biomedical and health data sources is given.*

Keywords: BIOMEDICAL DOMAIN, ONTOLOGY, DATA, INFORMATION, INTEGRATION, LINKED OPEN DATA, OWL, RDF

1. Introduction

The biomedical domain is distinguished by rapidly and versatile implementation of the achievements of ICT. Significant and with continuous rates increases the amount of accumulated and used information which increasingly is stored in a different size, complexity, levels of abstraction, perspectives and areas of application databases and lexical glossaries, directories and ontologies. Particularly strong growth marks represention of biomedical entities, their terms and relations in form of vocabularies, terminologies and ontologies. Some of them contain overlapping information and some application may require a domain ontology which spans several ontologies. "Ontology integration" consists in establishing relations between concepts belonging to different ontologies. The effective use of this information stored on different sources, in different forms and formats is essential.

Biomedical ontologies provide essential domain knowledge to drive data integration, information retrieval, data annotation, natural-language processing and decision support. The main aim of the proposed paper is to suggest an ontology based approach supporting the data and information integration in the biomedical domain based on the concept of Linked Open Data supported by Linked Open Vocabularies (LOV) [http://lov.okfn.org/dataset/lov] and the OWL version of schema.org namespace [https://schema.org/docs/documentation/html], enhanced with the D2RQ platform for RDB2RDF transformation.

The paper is organized in 5 parts. After the introduction, in part 2 a short overview of existing biomedical ontologies, classified in four basic categories, is given. The third part of the paper discusses the approaches and problems by data and information integration. In Part 4 the suggested approach for ontology based data and information integration is described. Applicability of the suggested approach is illustrated with a case study in part 4. Finally some conclusions are made.

2. Short overview of biomedical ontologies

Too many research efforts have been made in the biomedical domain for creation and use of different in type and size ontologies. Most of current biomedical ontologies are principally taxonomic hierarchies with sparse relationships. Four basic categories of ontologies are in use in the biomedical domain, as shown in Fig.1: top-level ontologies or upper-level, upper-level domain ontologies, domain ontologies and application ontologies.

The top-level ontologies are one of the main pillars used as a formal foundation for building domain ontologies. The primary purpose of top-level ontologies is to describe the general concepts through a framework of axioms and definitions or categories such as continuant, process and boundary and relationships such as "is_a" (for subtype) and "part_of" [1, 2]. The most popular top-level ontologies, useful for the biomedical domain are DOLCE (Descriptional Ontology for Linguistic and Cognitive Engineering) [3] and BFO (Basic Formal Ontology) [4]. DOLCE is a high-level, domain-independent conceptual framework for representing meaning. BFO is a foundational ontology that aims to adopt the structural vocabulary introduced for the characterization of DOLCE as it concerns universals, without fussing about the modal interpretation [3, 1]. The next category includes ontologies, which contain core concepts of a given domain and is working as an interface between the top-level and different domain categories. Some of the most popular representatives of this category are UMLS (Unified Medical Language System) [5] and GALEN (General Architecture for Languages, Enclopedias and Nomenclatures in Medicine) [6]. UMLS support the development of computer systems in biomedical domain, based on the UMLS Knowledge Sources, composed of Metathesaurus, containing information about biomedical and health-related concepts, their various names, and the relationships among them; Semantic Network, providing consistent categorization of all concepts represented in the Metathesaurus and the SPECIALIST Lexicon including many biomedical terms in English. GALEN is one of the first attempts for representing coded patient information and uses common reference model for representing medical concepts.

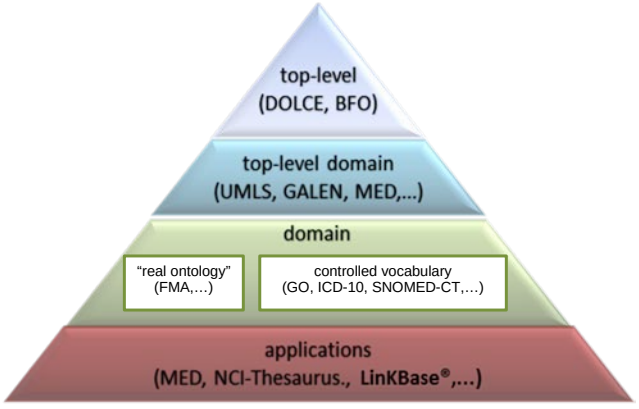

Fig.1*: Biomedical ontologies classification*

The largest category of ontologies in biomedical domain is this of domain ontologies, which represent knowledge about particular part of the world in a way that is independent from specific objectives, through a theory of the domains [1]. In reality, the number of real ontologies is not large. Therefore, they are considered two subgroups: real ontologies and controlled vocabularies. Representative of the first subgroup is the Foundational Model of Anatomy (FMA) [7, 8] that represents

declarative knowledge about structural organization of the human body with the intent of expanding the anatomical content of UMLS. The next subgroup includes the so called "controlled vocabularies" which are purpose-oriented and designed to meet particular needs, such as annotating biological databases (Gene Ontology (GO) [9, http://www.geneontology.org] and other OBO (Open Biological and Biomedical ontologies) ontologies [http://www.obofoundry.org/]) or medical records (ICD-10 [http://www.who.int/classifications/icd/en/], SNOMED CT [http://www.ihtsdo.org/snomed-ct]). OBO are set of orthogonal interoperable reference ontologies for biomedical domain built in the frame of OBO Foundry based on BFO, such as GO. GO as a part of OBO provides structured, controlled biological terminology that describes gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Medical records are presented by ICD-10 that is the 10th release of International classification of diseases, disorders, injuries and other related health conditions, and is used in clinical care, research, health care and globally for statistics and trends. The ICD-10 defines the universe of diseases, disorders, injuries and other related health conditions. One of the most comprehensive, semantically accurate, multilingual clinical healthcare terminologies in the world is represented in SNOMED CT, developed by IHTSDO (International Health Terminology Standards Development Organisation) and is used across all health systems, services and products in the world. The ontology concepts represent terms and processes that capture the meanings associated with healthcare related observations, procedures, functions and therapies. The NCI (National Cancer Institute) Thesaurus [http://bioportal.bioontology.org/ontologies/NCIT] is a description of logic-based terminology for clinical care, translational and basic research, public information and administrative activities, and is available on the NCI Term Browser [https://nciterms.nci.nih.gov/ncitbrowser/pages/].

The application ontologies, describes the semantic of a single information resource and are useful for terminology – oriented applications. For example the Medical Entities Dictionary (MED) [http://med.dmi.columbia.edu/] is a concept-oriented metadata dictionary in use at New York Presbyterian Hospital (NYPH) and is a repository of medical terms arranged in a semantic network. The concepts contained in MED include those from ICD-9CM, UMLS and LOINC. Another application ontology, designed to integrate terminologies and databases with applications in natural language processing and information retrieval, is LinKBase® [10].

3. Ontology based data and information integration

As discussed in the above part of the paper, there exist a lot of different specialized biomedical ontologies, databases, information systems, applications, semantic nets that combine data from several sources, each of which is accessed through an API specific to the data provider. The existence of a specialized API for each data set creates a landscape where significant effort is required to integrate each novel data set. Consequently, data returned from Web APIs typically exists as isolated fragments, lacking reliable onward links signposting the way to related data [11]. The applied approaches for data and information integration based on reference models and semantic nets fail to achieve a high degree of integration.

Linking biomedical data distributed in different models and representations requires a standard mechanism for specifying the existence and meaning of connections between items described in this data. The concept Linked Data is introduced by Tim Berners-Lee as unique identification and links between heterogeneous web resources to detect and retrieve information. Linked Data Platform 1.0 (W3C Recommendation since 26 February 2015) defines a set of rules for HTTP operations on web resources to provide an architecture for read-write Linked Data on the web. Linked Data Platform combines a common data model, a standard mechanism to access the data using the HTTP protocol, HTML hyperlinks and approved shared domain vocabularies. The success of Linked Open Vocabularies (LOV) as a central information point about vocabularies is symptomatic of a need for an authoritative reference point to aid the encoding and publication of data. The definitions of terms (classes, properties, or instances) provided by the LOV vocabularies bring clear semantics to descriptions and links thanks to the formal language they use, providing the semantic glue enabling data to become meaningful data. Vocabulary terms are identified by public URIs and can be linked inside a vocabulary and across vocabularies. The latest version of the Ontology for Biomedical Investigations (OBO) in LOV is from 01.08.2015. The large and growing set of terms in the schema.org namespace includes (and references) many established terms. Health and medical types in the schema.org ("MedicalEntity" and subtypes) are useful for content publishers that wish to mark up health and medical content on the web.

A key factor in the reusability of data is the extent to which it is well structured. Ontologies are succeeding to a large degree as a knowledge representation and data integration. Ontology matching is a solution to the semantic heterogeneity problem. It finds correspondences between semantically related entities of ontologies. In [12] ontology matching is formalized based on a unified account over a lot of previous works. The matching operation determines an alignment A′ for a pair of ontologies O_1 and O_2 as shown in Fig.2.

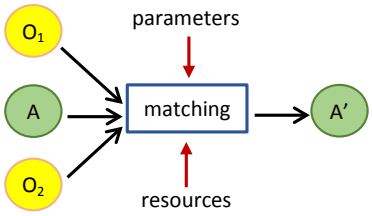

Fig.2: *Ontology matching*

An alignment is a set of correspondences between entities belonging to the matched ontologies. Alignments can be of various cardinalities: 1:1 (one-to-one), 1:m (one-to-many), n:1 (many-to-one) or n:m (many-to-many). Given two ontologies, a correspondence is a 4-uple <id, e_1, e_2, r>, such that: id is an identifier for the given correspondence; e1 and e_2 are entities, e.g., classes and properties of the first and the second ontology, respectively; r is a relation, e.g., equivalence (=), more general (⊒), disjointness (⊥), holding between e1 and e_2. The correspondence <id, e_1, e_2, r> asserts that the relation r holds between the ontology entities e1 and e_2. Correspondences have some associated metadata, such as the correspondence author name. A frequently used metadata element is a confidence in the correspondence (typically in the [0, 1] range).

4. Description of the suggested approach for data and information integration

The approach suggested in this paper in order to represent and integrate different biomedical data in RDF employing the standardized and widely used vocabularies as well as the publishing term definitions via the Linked Data principles.

4.1. Linked Open Data

The Linked Data approach offers significant advantages over current practices for creating and delivering biomedical data. Linked Data and especially Linked Open Data is sharable, extensible, and easily re-usable [13]. It supports multilingual functionality for data and user services, such as the labeling of concepts identified by language-agnostic URIs. Linked Data is expressed using standards such as RDFS, SKOS, and OWL, which specifies relationships, integrate information from multiple sources and are able to infer new information from a set of asserted facts. Quite often, the RDF/OWL ontologies might be automatically generated from legacy data sources such as spreadsheets, XML files and databases.

4.2. D2RQ open source software platform

The approach does not exclude the legacy databases and traditional Web content, because RDF serializations can be generated on-the-fly and present data from multiple datasets as if it were within a single database.

The D2RQ [http://d2rq.org/] is open source software platform for accessing relational databases as virtual, read-only RDF graphs. It offers RDF-based access to the content of relational databases without having to replicate it into an RDF store. D2RQ platform consists of:

- D2RQ Mapping Language - declarative XML-based language for describing the mapping between the relational schema and OWL / RDFS ontologies;
- D2RQ Engine - plug-in in Jena Framework (Java environment, providing software tools and libraries for the Semantic Web and Linked Data), which uses D2RQ Mapping Language to rewrite queries in SQL to the database;
- D2R Server - HTTP server that allows HTML links to data and SPARQL endpoint to query the data.

D2RQ Mapping Language creates a mapping file, analyzing the database scheme as each table is transformed into a new RDF class with the same name and each field is transformed into property data type. D2RQ Mapping Language also creates Uniform Resource Identifiers (URIs) of the database data. D2RQ platform can be used for: (i) query a non-RDF database using SPARQL, (ii) access the content of the database as Linked Data over the Web, (iii) create custom dumps of the database in RDF formats for loading into an RDF store, (iv) access information in a non-RDF database using the Apache Jena API.

4.3. OWL ontology of schema.org

The scope of medical terms in schema.org is broad, and is intended to cover both consumer- and professionally-targeted health and medical web content; as a result, any particular piece of content is likely to use only a subset of the schema. “MedicalEntity” in schema.org is not intended to define or codify a new controlled medical vocabulary, but instead to complement existing vocabularies and ontologies. As a schema, its focus is on surfacing the existence of and relationships between entities described in content. The schema does provide a way to annotate entities with codes that refer to existing controlled medical vocabularies (such as MeSH, SNOMED, ICD, RxNorm, UMLS, etc).

The class “MedicalEntity”, the most generic type of entity related to health and the practice of medicine is presented in Fig.3. The owl version of “MedicalEntity” class consists of 17 subclases:

- “AnatomicalStructure” - consists of subclasses defined for parts of the human body as components of an anatomical system: organs, tissues, and cells;
- “AnatomicalSystem” – is used to describe groups of anatomical structures that work together to perform a certain task such as organ systems: circulatory, digestive, endocrine, integumentary, immune, lymphatic, muscular, nervous and other systems;
- “MedicalCause” - includes cardiovascular, chemical, dermatologic, endocrine, environmental or gastroenterological causes, etc.;
- “MedicalCondition” - includes diseases, injuries, disabilities, disorders, syndromes, etc.;
- “MedicalContraindication” - is a condition or factor that serves as a reason to withhold a certain medical therapy;
- MedicalDevice” - is "Any object used in a medical capacity, such as to diagnose or treat a patient";
- “MedicalGuideline” - are recommendations made by a standard society (e.g. ACC/AHA) or consensus statement that denotes how to diagnose and treat a particular condition;
- “MedicalIndication” is a condition or factor that indicates use of a medical therapy, including signs, symptoms, risk factors, anatomical states, etc.;
- “MedicalIntangible” - is a utility class that serves as the umbrella for a number of 'intangible' things in the medical space;
- “MedicalProcedure” - is a process of care used in either a diagnostic, therapeutic, or palliative capacity that relies on invasive (surgical), non-invasive, or percutaneous techniques;
- “MedicalRiskEstimator” - is defined as "Any rule set or interactive tool for estimating the risk of developing a complication or condition";
- “MedicalRiskFactor” - is anything that increases a person's likelihood of developing or contracting a disease, medical condition, or complication;
- “MedicalSignOrSymptom” - is any indication of the existence of a medical condition or disease;
- “MedicalStudy” - is an umbrella type covering all kinds of research studies relating to human medicine or health, including observational studies and interventional trials and registries, randomized, controlled or not;
- “MedicalTest” - is any test, typically performed for diagnostic purposes;
- “MedicalTherapy” - is defined as medical intervention designed to prevent, treat, and cure human diseases and medical conditions, including both curative and palliative therapies;
- “SuperficialAnatomy” - consists of anatomical features that can be observed by sight (without dissection), including the form and proportions of the human body as well as surface landmarks that correspond to deeper subcutaneous structures.
- “MedicineSystem“ - is an enumeration class for systems of medical practice: “Western conventional”, Homeopathic, Osteopathic “Traditional Chinese”, etc. The property “MedicalCode” for the “MedicalEntity” can be taken from a controlled vocabulary or ontology such as ICD-9, DiseasesDB, MeSH, SNOMED-CT, RxNorm, etc.

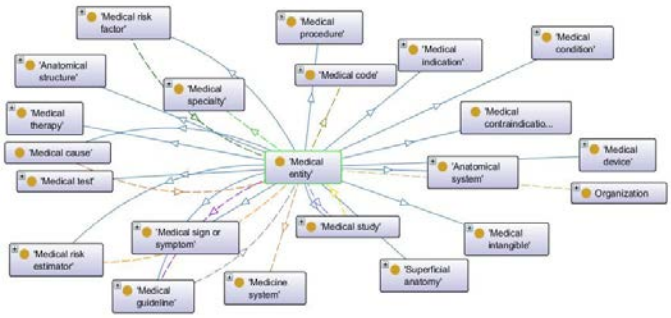

Fig.3: *The OWL version of “MedicalEntity”*

The techniques and platform described above are combined as illustrated in Fig.4 in order to organize a semantic virtual warehouse, allowing the integration and sharing of different data and information from various resources in biomedical domain.

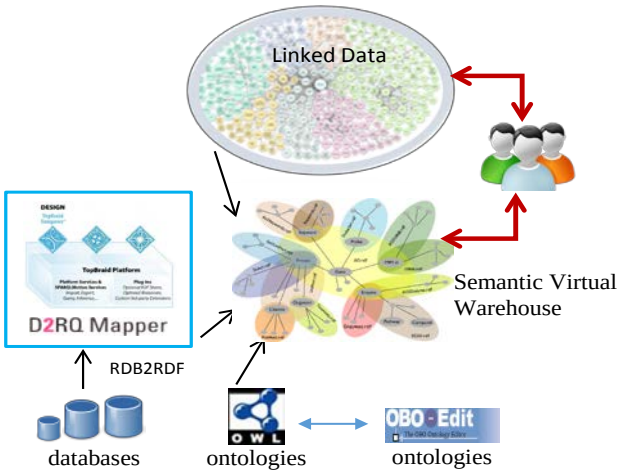

Fig.4: *Illustration of the suggested approach*

4. Applicability of the suggested approach

The proposed approach will be illustrated through the integration of ontology based system for managing clinical data to the owl version of schema.org (updated 2015-08-25). D2RQ platform is used for database to ontology mapping. The database "patient.sql" downloaded from http://sourceforge.net/ is used by "OpenPatientOS" - an information system for managing patient records. The system manages patient personal, medical, and billing records through an easy to use Swing user interface. The "OpenPatientOS" offers personal data entry and management, patient medical records management, patient billing management system, reports creation and user's management system. The system is intended to be used by administrators and physicians; we aim to extend it to patients so that the patients are able to look for their own medical records. The ER model of database and the corresponding ontology model derived with D2RQ mapping language are presented on Fig.5 and Fig.6, respectively.

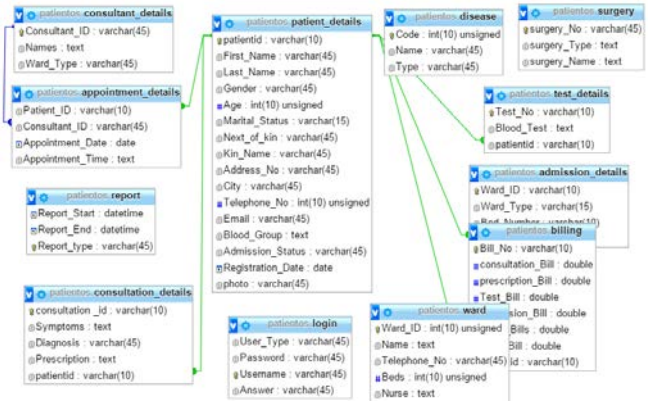

Fig.5: *ER data model of "patient.sql"*

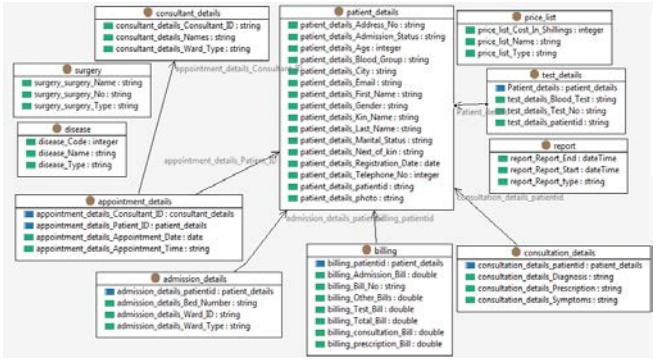

Fig.6: Ontology model corresponding to "patient.sql"

The patient ontology is integrated with the owl version of schema.org ontology in IDE Topbraid Composer [http://www.topquadrant.com/tools/IDE-topbraid-composer-maestro-edition/]. The class "patient:test_details" is declared as a subclass of the class "schema:BloodTest" from the schema.org ontology (Fig.7). As a result of the ontology matching <id31, schema:BloodTest, patient:test_details, ⊂ >, the individuals of class "patient:test_details" inherit 16 additional properties from the schema.org ontology. While the database properties for the class "test_details" are olny "patient:Blood_Test", "patient:Test_No", and "patient:test_detais_patienid", the additional properties: "schema:affectedBy", "schema:signDetected", "schema:usedDevice", "schema:medicineSystem" and "schema:usedToDiagnose" extend the information for a particular patient. A trail SPARQL query for Information retrieval from the integrated model is shown on Fig.8.

5. Conclusions

Based on the analysis of the current state of development of ontologies in biomedical applications and the approaches for data and information integration, discussed in part 2 and part 3 of the paper that most of the data and information resources do not conformed to the formal principle of ontology design and therefore cannot be reused for other purposes and applications and does not support automatic reasoning. The approaches, used to integrate heterogeneous data and information, as reference models and biomedical semantic nets fail to achieve a high degree of integration. The suggested approach allows integration of different heterogeneous data and information resources: Linked Open Data vocabularies, ontologies and relational databases.

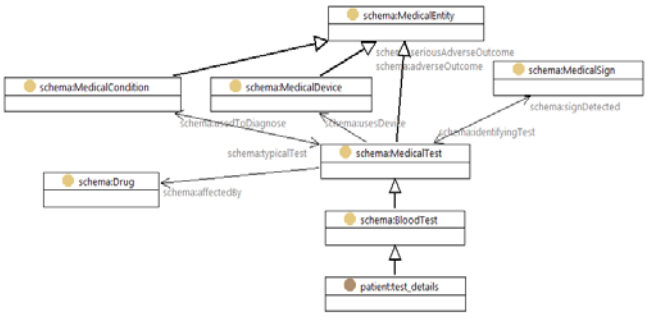

Fig.7: *Integration results*

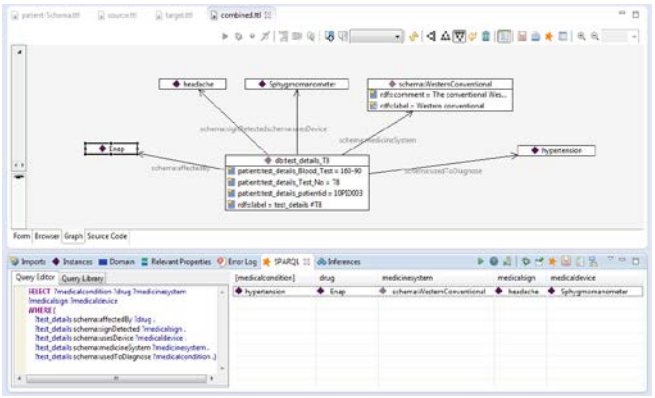

Fig.8*: Information retrieval from the integrated model*

6. References

1. Rosse C., Kumar A., Mejino J. L.V, Cook D. L., Detwiler L. T., Smith B., A Strategy for Improving and Integrating Biomedical Ontologies, AMIA Annual Symposium Proc. 2005, pp.639–643.
2. Alexander C. Y., Methods in biomedical ontology, Journal of Biomedical Informatics, Vol.39, Issue 3, June 2006, pp.252–266.
3. Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., Ontology library, Deliverable D18 of the IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web, 2003.
4. Basic Formal Ontology (BFO), http://ifomis.uni-saarland.de/bfo/.
5. U. S. National Library of Medicine, Unified Medical Language System (UMLS), UMLS® Reference Manual, 2015, http://www.ncbi.nlm.nih.gov/books/NBK9675/.
6. Cimino J. J., Zhu X., The practical impact of ontologies on Biomedical Informatics, IMIA Yearbook of Medical Informatics, 2006, pp.1-12.
7. Rosse C., Mejino J. L. V., A reference ontology for biomedical informatics: the Foundational Model of Anatomy, Journal of Biomedical Informatics, 2003, 36(6), pp.478–500.
8. Foundational Model of Anatomy, University of Washington, 2015 http://sig.biostr.washington.edu/projects/fm/AboutFM.html.
9. Gene Ontology Consortium, The Gene Ontology (GO) project in 2006. Nucl. Acids Res. (Database issue) 2006, 34, pp.322–326.
10. Simon, J., Fielding, J., Santos, M.D. & Smith, B. (2004). Reference ontologies for biomedical ontology integration and natural language processing. In *Proceedings of EuroMISE 2004*, Prague.
11. Heath T., Bizer C., Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool, 2011.
12. Shvaiko P., Euzenat J., Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering, Institute of Electrical and Electronics Engineers (IEEE), 2013, 25 (1), pp.158-176.
13. Library Linked Data Incubator Group Final Report, W3C Incubator Group Report 25 October 2011.