Using Semantic Web Technologies to Power LungMAP, a Molecular Data Repository

M. C. Krzyzanowski
RTI International
Research Triangle Park, NC
United States
mkrzyzanowski@rti.org

J. Levy
RTI International
Research Triangle Park, NC
United States
josh.levy@rti.org

G. P. Page
RTI International
Research Triangle Park, NC
United States
gpage@rti.org

N. C. Gaddis
RTI International
Research Triangle Park, NC
United States
ngaddis@rti.org

R. F. Clark
RTI International
Research Triangle Park, NC
United States
rclark@rti.org

ABSTRACT

As scientific research evolves, data continue to grow at an exponential rate. This growth calls for a need for more data repositories to store the data, and the creation of additional centralized repositories to provide standards for researchers. Common data repositories allow for collaboration and easier sharing of data, critical for further advancement of scientific understanding of a variety of topics. LungMAP (the Molecular Atlas of Lung Development) is an open-access reference resource that provides a comprehensive molecular atlas of the normal developing lung in humans and mice and provides data and reagents to the research community. The database utilizes RDF, SPARQL, and OWL. LungMAP exemplifies the use of semantic web technologies to provide a collaborative and open access data application for the scientific research community.

CCS CONCEPTS

• **Information systems~Graph-based database models** • **Information systems~Resource Description Framework (RDF)** • **Information systems~Ontologies**

SBD'17, May 19, 2017, Chicago, IL, USA

DOI: http://dx.doi.org/10.1145/3066911.3066916

KEYWORDS

RDF, OWL, Semantics; Ontology, Cloud Computing, OpenLayers

ACM Reference format:

M. C. Krzyzanowski, J. Levy, G. P. Page, N. C. Gaddis, and R. F. Clark. 2017. Using Semantic Web Technologies to Power LungMAP, a Molecular Data Repository. In *Proceedings of the International Workshop on Semantic Big Data (SBD'17), Chicago, IL, USA, May 19, 2017,* 6 pages.
DOI: 10.1145/3066911.3066916

1 INTRODUCTION

The National Heart, Lung, and Blood Institute (NHLBI) conducts and supports research on heart, lung and blood diseases and promotes a collaborative environment in partnership with patients, health care professionals, scientists and the community [1]. Genetic and environmental abnormalities can affect early fetal lung development, and result in abnormally developed lung structure and function. These abnormalities may lead to susceptibility for several childhood and adult lung diseases, such as COPD, cystic fibrosis (CF), and asthma [2]. Characterizing the array of active or inactive genes, peptides, metabolites, methylation, and miRNA at different time points during development may yield increased understanding of the role of genes, proteins, and their interactions on lung development and diseases. Currently, there are several reported susceptibility candidate genes in mice where change in gene expression can alter proper development of the lungs and related structures, and ultimately result in the onset of mild to severe lung-related diseases early to late in life [2]. However, the ability to identify genes, gene expression and protein function throughout development in mice and human will yield a comprehensive dataset of candidates for further research endeavors.

Any large-scale biological research initiative benefits with contributions from multiple teams in order to build a complete picture of what events are occurring and for peer-reviewed

confirmation of the data generated. To fulfill the need for better understanding of key molecular markers and their expression patterns throughout development, the NHLBI launched the Molecular Atlas of Lung Development Program (LungMAP), a molecular atlas of the normal developing lung from human and mouse. The goal of this resource is to serve as a reference for the research community and promote better understanding of molecular lung development. At hand was the challenge of coordinating and integrating the experimental data from LungMAP Consortium members (the Human Tissue Core (HTC) and the Research Centers (RCs)), and creation of novel tools to analyze the experimental data.

The LungMAP web application (www.lungmap.net) is structured around a high-resolution ontology incorporating well-characterized structural and functional anatomic components, distinct histological tissue compartments, and generic and specific cell types defined at the molecular level. The web application features a wide breadth of molecular data, including images, videos, time-course data, RNA, lipids, metabolites, and protein data. The use of ontologies and triple store databases allows for integration of diverse data types with any sematic content generated by current LungMAP Consortium members and ultimately by investigators outside the Consortium. This application facilitates data storage at one centralized location for multiple research groups and promotes collaboration among Consortium members and other interested parties, further adding to the depth of the data and evolution of the application.

2 METHODS

2.1 Data Architecture

The overall basic data architecture for LungMAP is shown in Fig. 1. Data is stored in a triple store database (OpenLink Virtuoso), named Bioinformatics REsource ATlas for the Healthy lung (BREATH) database. Data integration is mapped to the Resource Description Framework (RDF) data model of graphs, following *subject, predicate, object* triples and W3C Web Ontology Language (OWL), a computational logic based language. RDF provides the needed flexibility of modeling the diverse datasets being uploaded from multiple HTC and RCs. The triple store database makes it possible to extend the ontology without needing to redo the data structure.

2.2. Ontologies

The ontologies powering LungMAP are broken up into four categories: (1) anatomical ontology for human lung maturation, (2) anatomical ontology for mouse lung maturation, (3) cell ontology for human lung maturation and (4) cell ontology for mouse lung maturation. There are separate ontologies for human and mouse tissues. These ontologies are continually evolving, guided by new discoveries made by the LungMAP Consortium. Importantly, these ontologies direct the structure and search functions of LungMAP.

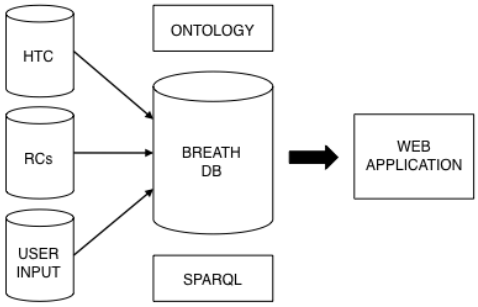

Figure 1: Diagram of the data architecture and semantic web technologies underlying LungMAP web application. Human Tissue Core (HTC); Research Centers (RCs); Database (DB).

The LungMAP human anatomy (LMHA) or LungMAP mouse anatomy (LMMA) ID begins with the 4-digit abbreviation specific to the species followed by a 10-digit number. The triple store allows transitivity. For example, if B is a subclass of A and C is a subclass of B, then C is also a subclass of A. Applying this principle to a partial example, in the LungMAP ontology (Fig. 2), the class "trachea" contains three subclasses: "trachea bifurcation", "trachea lumen" and "trachea wall". "Trachea wall" contains two subclasses, "tracheal mucosa" and "tracheal lamina propria", both of which by association, are also subclasses of the parent class "trachea". This allows retrieval of information in anatomical clusters that are linked to the top class of trachea or a subclass, such as trachea epithelium.

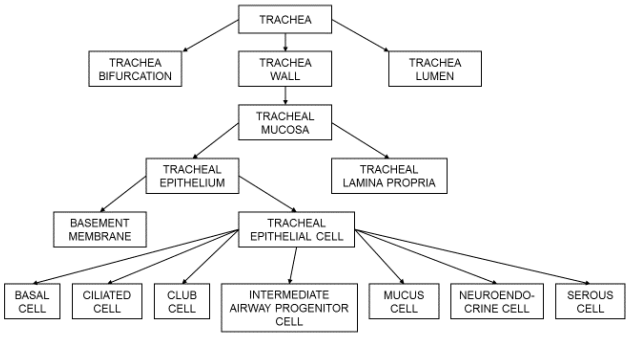

Figure 2: Partial example anatomical ontology for LungMAP.

Furthermore, each term in the ontology contains a unique identifier (such as LMHA0000000142), a name ("ciliated cell") and a description ("A columnar epithelial cell with microscopic hair-like processes, i.e., motile cilia, that extend from the apical cell surface. These cells are found in the epithelia of the conducting airways (trachea, bronchi, and bronchioles), where they sweep mucous, dirt, dust, bacteria and other pathogens up and out of the lung, bronchi, and trachea."). Unlike annotations, these LungMAP terms are formalized based on known anatomy of the developing mouse and human lungs, and provide the logic to query an information set and analyze across sets. For example, a query on anatomy for "ciliated cell" can use the hierarchical

relationships to identify related anatomical structures, such as "basal cell", "club cell" and "tracheal epithelial cell." From there, further information can be gleaned on these related anatomical structures.

2.3 LungMAP Web Portal and Features

As part of the data coordinating center (DCC), our team is tasked with maintenance of the LungMAP portal and BREATH. We built the centralized data repository and public interface for LungMAP through the creation and management of:

- standard operating procedures for data management
- existing lung development results
- ontologies for lung development, structures, and cross species comparison
- experimental data from RCs and biologic sample data from the HTC
- novel tools to analyze the experimental data

The integration of the experimental and biologic sample data is visualized at the LungMAP web portal (Fig. 3), which has been designed with a number of use case scenarios, including:

- A researcher interested in browsing available data of a particular experiment type.
- A researcher interested in finding data from all experiment types related to a specific term of interest.
- A researcher seeking specific reagents or detecting certain genes or proteins during lung development.

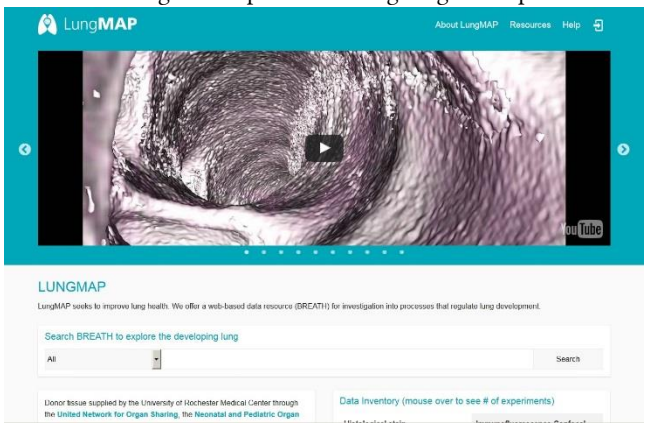

Figure 3: The landing page for LungMAP web portal.

2.3.1 Basic search. The portal provides a number of search methods to cover integrated content. A user may limit their search to a specific category (gene, lipid, etc.) or search for all results that match their search term. A search for all results will bring up results broken down by predefined categories, and allows the user to browse through the results.

2.3.2 Information on entities of interest. LungMAP provides overview pages for entities such as genes, proteins, anatomical structures, and developmental time points. Each page illustrates the integrated data from the Consortium HTC and RCs. These pages can be reached either by search of a general term or an entity. For example, searching by a gene's entity ID will bring up a results page including the gene name and provide additional information about the specific gene, such as species and genomic location. Upon searching for the term "acta1", a results page containing the same result from a search for "Gene_ID_11459" as well as other matches for "acta1" would be retrieved.

2.3.3 Image annotation. If a user is signed in and has annotator privileges, any image accessed through the image annotation page can be annotated. Annotations are powered by OpenLayers [3], an open source JavaScript library used to create dynamic maps. The user can pick a point or arrow, or highlight an area using a polygon tool, and then add terms to the specified annotation. New annotations are reviewed by the Consortium, and upon approval, are available to be viewed publicly by any visitor to the application.

2.3.4 Data and documents download and upload. Registered users have access to documentation and preliminary shared data. Documentation includes Consortium meeting agendas, presentations, and instructions for user only features. Files are available for download.

3 RESULTS

3.1 Example A: User to Ontology to Application

As a case study, a user wants to find more information about the mouse gene Acta1. Upon reaching the individual gene entity page, the user is curious to see what type of single-cell RNA-seq experiments have results which include the Acta1 gene (Fig. 4). Clicking the sidebar menu link, a query transverses a path to retrieve all single-cell RNA-seq experiments that have results for Acta1. The SPARQL query is set up in the following manner:

- Prefixes to save on typing
- Definition of the columns to select
- Graphs to use for the query
- Triple patterns to be matched

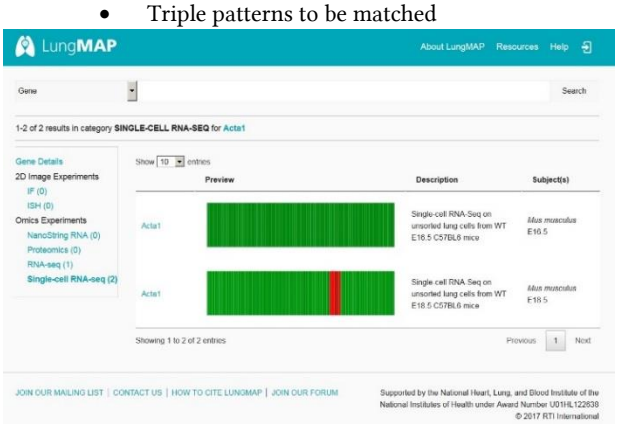

Figure 4: Use case example of the web interface rendering of a SPARQL query results.

Table 1: List of prefixes used in examples A and B.

Prefix	Location
owl	http://www.w3.org/2002/07/
owl2	http://www.w3.org/2002/07/owl#
mont	http://ontology.lungmap.net/ontologies/mouse_anatomy#
lm	http://ontology.lungmap.net/ontologies/expression_ontology#

As seen in Fig. 5, the query specifies the triples in the context of an ontology. The results must satisfy all the specified conditions, given the input of an Entrez ID and experiment type ID. The input of experiment type ID allows for information to be gathered about the experiment ID, analysis ID and overall expression. The analysis result (part of the analysis ID) is mapped to the entity ID and the entity ID is mapped to the Entrez ID, which is one of the initial input values. Through these mapping events, we are able to retrieve the Gene ID and taxon ID and organism label. As a result, only two results fit these triple patterns and are rendered in the web application, as seen in Fig. 4.

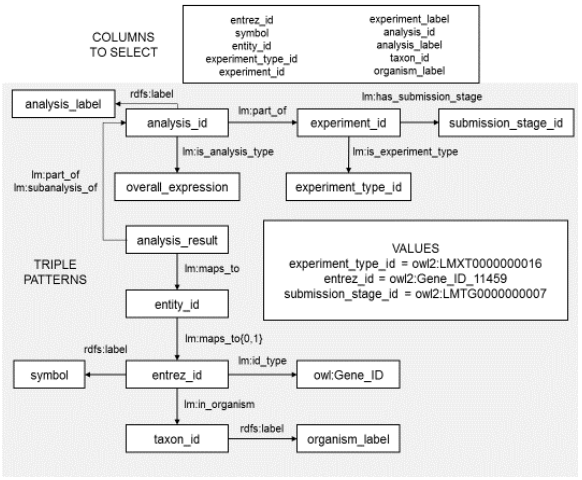

Figure 5: Triple patterns conducted during use case example A.

The data retrieved then helps build the result table and further proceeds to conduct additional queries to retrieve the data to build the heat maps (Fig. 4). Briefly, the mouse gene specific entity ID and the analysis ID from the columns selected in Fig. 5 are sent to conduct a query that specifies triples to obtain the z scores and cell types that build a heat map using the HighCharts javascript library [13]. Overall, a range of data and triples are used in order to obtain the final visual result as rendered on the web application.

3.2 Example B: Integrating Triple Store and OpenLayers

OpenLayers is a mapping JavaScript library commonly used for mapping of geographic locations. LungMAP utilized this technology in another way – to map anatomical features on immunofluorescence-confocal and histological stain images of lung tissue. In short, the markers at specified "map coordinates" are annotated features, such as cells or structures in an image relative to the entire image itself (Fig. 6). The markers linked to each feature are stored in the BREATH database, and are rendered on the image upon loading of the page.

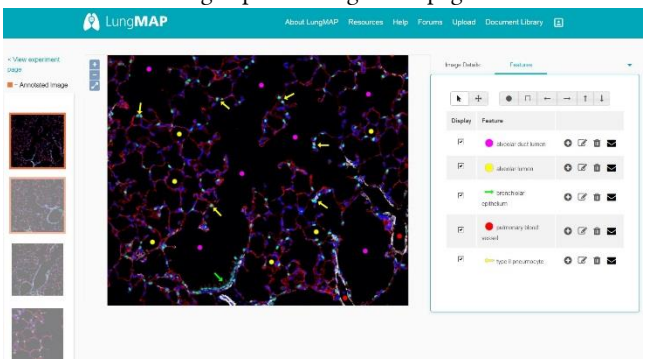

Figure 6: An immunofluorescence-confocal image annotated with anatomical features.

In example A, a SPARQL query specifies the triples in the context of an ontology. The results must satisfy all the specified required conditions, given the input of an image ID (Fig. 7). The input of image ID allows for information to be gathered about the feature ID and its associated coordinate string, color, rotation shape type and comment. Optionally, the annotation ID may be a part of the feature ID, and retrieves the associated condition, condition label, image feature annotation and status. In this example, 27 image features are returned, and rendered on the image through OpenLayers and in a right column, each with the option to edit the feature.

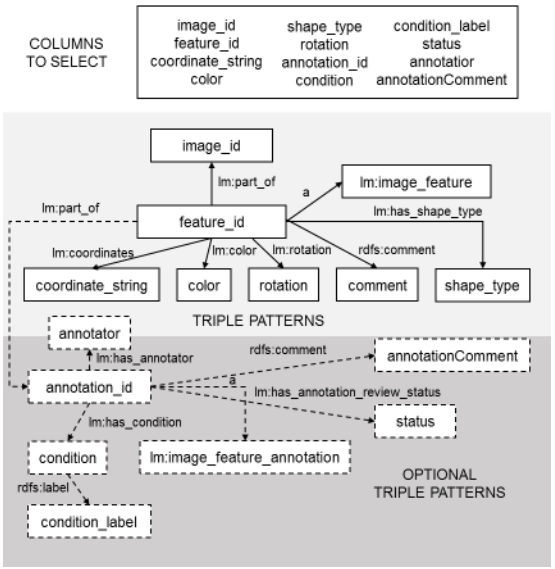

Figure 7: Triple patterns conducted during use case example B. Triples that must be fulfilled are with a solid outline. Additional optional triples are indicated by a dashed outline.

3.3 LungMAP Relies on and Promotes Collaboration

LungMAP as an application ultimately relies on the contributions of Consortium members, HTC and RCs. Large datasets are uploaded through SFTP from 10 member academic research institutions and hospitals, and data is deposited into the BREATH database by the Data Coordinating Center at RTI International [4]. This allows data from different studies to be merged into one large searchable dataset, facilitated by RDF entities and defined by LungMAP's ontologies. In addition to large datasets, registered users can upload preliminary data, images, videos, publications, and presentations through the web application file upload feature. LungMAP also features a user forum powered by Discourse, an open source discussion platform [5]. The forum allows registered users to discuss matters related to the LungMAP data and web application, and provides the environment to promote discussion and potential collaboration among the researchers committed to the understanding of lung development.

The LungMAP Consortium has an open access data policy, and thus any visitor to the web application can search, download, and view data available through the LungMAP website. All visitors can sign up for a user account, and comment on existing data or contact the Consortium, contributing to curation and accuracy of the data presented in the application. For example, if a registered user with annotator privileges were to search for "alveolar" and head to the image annotation page, the user could view non-approved annotations, promoting review of the data. Users with annotator account privileges can add annotations to experiment images for approval and request new terms to label the annotations, further contributing to the curation of the existing data, the molecular mapping of lung development and the evolution of the ontologies powering LungMAP. Importantly, the open access data policy builds a level of trust among the contributors and validity of the data presented.

Finally, by linking the authors of the datasets, annotations and other files that are available on the web application, the chances of discussion and collaboration are increased. Studies suggest that collaboration has a significant impact on scientific productivity and impact [6] and collaborations in biology have increased over the years [7]. Exchange of research materials, divided labor and finding collaborators with key backgrounds only helps make large-scale projects, such as the molecular atlas goals of LungMAP, feasible, and increases the speed at which data is produced and reviewed.

4 DISCUSSION

As scientific research continues to generate large data sets, it encompasses a need for development of central applications and data standards to promote collaboration among different research institutions. The National Institute of Health established the Big Data to Knowledge initiative in 2012 to promote development of tools and to increase data accessibility to advance understanding of human health and disease [8]. These initiatives recognize the need to establish data repositories and common standards to accommodate the exponential growth of data collected, and the shift toward digitalization to store this information.

With the success in the generation of large data sets in a high-throughput manner, experimental technologies now produce complex data sets that require advances in managing and relating these datasets. In biology, it is common to form relationships that are hierarchical in nature, such as in anatomy and signal transduction pathways. As a result, there has been a push in recent years to represent this biological knowledge in the form of ontologies, called "bio-ontologies" [9]. While the definition of an ontology has been applied in biology for many years, it is only more recently been recognized as such. There are multiple bio-ontologies on the web that help describe biological, specifically phenotypic (observable characteristics) data. Gene Ontology (GO) [10] incorporates biological processes, molecular functions and cell components, and allow a user to identify all terms with a given protein and find more detailed information about a specific gene. Another is BioPortal, which provides information about biomedical ontologies [11].

A challenge facing bio-ontologies is handling complex areas of knowledge, especially when searching established biology databases that are not yet structured and the data not relational. As LungMAP needed to integrate both mouse and human anatomical terms and datasets, we opted to create organism-dependent ontologies, as there are many similarities and differences between mouse and human models, and it was crucial that these differences were accounted for to properly make analysis across datasets biologically sound. Furthermore, the anatomical and cell maturation for mouse and human were separated into two ontologies to prevent a large increase in the size of the ontology per organism, as multiple cell types could be related to anatomical structures, resulting in a complex ontology that could have been laborious to maintain.

For LungMAP as well as other bio-ontologies, the life science field is highly dynamic, and consequently, the ontologies are equally not static and will need to evolve to reflect the changes. As new domain knowledge is acquired, the collaborative effort of the entire LungMAP team provides will be critical in removing outdated information and incorporating new requirements. For all bio-ontologies, there will be a need in the future to incorporate algorithms or use enrichment tools [12] to survey current ontology-based mappings and avoid completely relying on manual mapping maintenance. It is also important to note that typical ontology correspondences (i.e. is_a, part_of) are not the only relationships used in evolving bio-ontologies, and mappings with domain-specific semantics will also need to be accounted for (such as LungMAP's "is_experiment_type"). This provides possible challenges in integrating data across multiple domains that could be related, but do not share, the same domain-specific semantics.

5 CONCLUSIONS

LungMAP has illustrated a method to integrate diverse datasets and use semantic web technologies to make the datasets available through a web-based application. By generating ontologies purposed to the anatomical data at hand, the application can further maximize search results and also leave room for evolution of the human and mouse anatomical ontologies going forward. LungMAP uses ontologies to focus on a key area of scientific research to allow for a streamlined and cross-experimental approach to analyzing datasets that normally could not be easily comparable. Furthermore, by integrating anatomy knowledge and gene expression, omics experiments and imaging datasets, LungMAP takes a focused yet broad approach in extending the reach of the ontology, in comparison to an overarching theme of some bio-ontologies (such as gene ontologies or a specific organism anatomy ontologies). A similar method of ontology generation could in theory be applied to any biological dataset, and provides an example of the power of using ontologies for scientific research purposes.

As data evolve, so will the features of LungMAP. Additional features on dock include increased ability for user feedback, direct ways of communicating with annotators or authors of datasets, comments on specific annotations, and general comments on a specific experiment image. The domain-specific and general ontologies will also evolve as new scientific discoveries are made, making the data presented by LungMAP highly dynamic. In conclusion, we present LungMAP as one model of integrated semantic web technologies for large scale and diverse experimental datasets and presentation of data in a searchable manner through a web application interface.

ACKNOWLEDGMENTS

This work is supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number U01HL122638.

REFERENCES

[1] "National Heart, Lung, and Blood Institute: The NHLBI Mission Statement," Last modified March 11, 2009. Accessed January 30, 2017. https://www.nhlbi.nih.gov/about/org/mission.

[2] W. Shi, S. Bellsci, and D. Warburton. 2007. Lung development and adult lung diseases. *Chest* 132, 2 (2007), 651-656.

[3] OpenLayers, 2017. Retrieved January 30, 2017, from OpenLayers: https://openlayers.org/.

[4] LungMAP: LungMAP Team, 2016. Retrieved January 30, 2017, from LungMAP: http://www.lungmap.net/about/lungmap-team/.

[5] Discourse, 2017. Retrieved January 30, 2017, from Discourse: http://www.discourse.org.

[6] L.S. Wuchty, B.F. Jones, and B. Uzzi. 2007. The increasing dominance of teams in production of knowledge. *Science* 316, 5827 (2007), 1036–1039. DOI: 10.1126/science.1136099.

[7] N. Vermeulen, J.N. Parker, and B. Panders. 2013. Understanding life together: A brief history of collaboration in biology. *Endeavour* 37, 3 (2013), 162–171. DOI: 10.1016/j.endeavour.2013.03.001.

[8] Data Science at NIH: BD2K, 2016. Retrieved January 30, 2017, from Data Science at NIH National Institute of Health: https://datascience.nih.gov/bd2k.

[9] Open Biological Ontologies, 2017. Retrieved March 20, 2017, from Open Biological Ontologies: http://www.obofoundry.org/.

[10] Gene Ontology Consortium, 2017. Retrieved March 20, 2017, from Gene Ontology Consortium: http://www.geneontology.org/.

[11] BioPortal, 2017. Retrieved March 20, 2017, from BioPortal: https://bioportal.bioontology.org/.

[12] P. Arnold and E. Rahm. 2014. Enriching ontology mappings with semantic relations. *Data & Knowledge Engineering* 93 (2014), 1–18. DOI: 10.1016/j.datak.2014.07.001

[13] Highcharts, 2017. Retrieved March 20, 2017, from Highcharts: https://www.highcharts.com/.