Towards Building a Semantic Repository of Bioinformatics Resources

Maria M. Nisheva-Pavlova[1, 2], Peter L. Stanchev[2, 3], Pavel I. Pavlov[1]

[1] Faculty of Mathematics and Informatics, Sofia University, Sofia, Bulgaria
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
[3] Kettering University, Flint, USA
marian@fmi.uni-sofia.bg, pstanche@kettering.edu,
pavlovp@fmi.uni-sofia.bg

Abstract. The paper presents a work in progress directed to the creation of a semantic digital repository of scholarly resources in the area of bioinformatics. A special attention has been paid to the design of a lightweight subject ontology that will be used for automated conceptual annotation of the dynamically entering new materials and as a knowledge source for intelligent search over the repository. Some issues concerning the implementation of the repository and a suitable tool for search and document retrieval over its resources are discussed as well.

Keywords: Digital Repository, Metadata, Ontology, Semantic Search, Knowledge Discovery, Bioinformatics

1 Introduction

The paper attentions in the process for establishing an improved service-oriented architecture (SOA) for interoperable and customizable access techniques for building semantic repositories with improved content searching. We propose a design of a lightweight subject ontology that will be used for automated conceptual annotation of the dynamically entering new materials and as a knowledge source for intelligent search over the repository.

The field of applying our repository is Bioinformatics. It is a rapidly growing area that creates and applies computationally intensive methods and software tools for automated analysis and interpretation of biological data. It is one of the fields of research and practical development which convincingly demonstrate the advantages of the use of various kinds of semantic technologies and especially ontologies and ontology mapping techniques (Hoehndorf, Schofield, & Gkoutos, 2015), (Huang, J., et al., 2010).

The rapid rates of modern bioinformatics research lead to the frequent emergence of new data, relevant scientific publications, and teaching materials. Any adequate project of a digital repository of bioinformatics resources should take into account this peculiarity and provide flexible tools for annotation (and, more general, for semantic enhancement) of dynamically coming heterogeneous resources as well as advanced tools for

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 7, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2017. ISSN: 1314-4006, eISSN: 2535-0366

intelligent search and information retrieval. The rest of the paper is organized as follows: Section 2 presents a brief overview of the proposed conceptual model and work methodology; Section 3 presents the project core ontology; the last section contains some concluding remarks.

2 Project Conceptual Model and Work Methodology

On the base of analysis of existing repositories we designed a typical academic digital repository provided with an advanced ***metadata module***, an ***annotator*** and a ***semantic search engine***. It has been under development in order to give up convenient access to various kinds of research and teaching materials in the field of Bioinformatics: experimental datasets, articles, presentations, lecture notes, etc. as well as to serve as a basis for experiments in the field of annotation and semantic enhancement of collections of heterogeneous resources.

The ***metadata module*** consists of two sections – an *annotation section* containing annotations of the resources preserved in the repository in terms of different metadata schemes and an *ontology section* containing a set of ontologies that can be used for annotation purposes.

An ***annotation*** is a specific description – a form of metadata attached to a specific resource (a dataset, a particular database field, a whole document or a particular section of document content). Annotation provides additional information (metadata) about an existing piece of data. Semantic annotation enriches the existing documents and data with a context that is further linked to the available formally described domain knowledge and makes it possible to process complex search queries and to receive results that are not explicitly related to the originally formulated queries. Ontologies are the only widely accepted form for the description and management of open, sharable, and reusable knowledge in a way, which allows automatic interpretation and inference. They provide semantic enhancement of data and types of resources suggesting controlled vocabularies for annotations.

Currently the ontology section contains a lightweight core ontology named BIO that has been used as the primary source of metadata for the stored resources. We plan to extend it with a set of freely available subject ontologies like a proper basic version of the Gene Ontology (Schuurman & Leszczynski, 2008), the MGED ontology (Whetzel, P., et al., 2006), etc. All these ontologies will be used by the annotator in building various types of descriptions (annotations) in accordance with several metadata standards. A further goal is to elaborate a methodology for automated analysis of a given set of resources (data, research or teaching materials) and selection of appropriate ontologies to be used as sources of conceptual knowledge for their semantic enhancement. An experimental software tool for automated selection of a set of suitable ontologies will be developed as well. It will be supplied with an additional module for mapping bioinformatics ontologies intended to assist the annotation process.

The purpose of the ***semantic search engine*** is to provide adequate access to the complete palette of resources stored in the repository. It supports several types of intelligent

search and information retrieval using the available annotations of resources and implementing well-known techniques of ontology-based augmentation and refinement (disambiguation) of the user queries. The user has options for sorting and grouping the search results in accordance with different criteria.

The implementation of the search engine is based on our previous experience and results in building digital repositories and semantic digital libraries in various domains (Nisheva-Pavlova & Pavlov, Building a Digital Library with Learning Materials, 2009), (Nisheva-Pavlova, Shukerov, & Pavlov, 2015) (Stanchev, Nisheva-Pavlova, & Geske, 2010). Modern semantic technologies will be used for creating annotations and ontology mapping purposes.

3 The Project Core Ontology

The core ontology named BIO is designed especially for the discussed project. It consists of seven interrelated sub-ontologies without formal hierarchical relations each with another (see fig. 1).

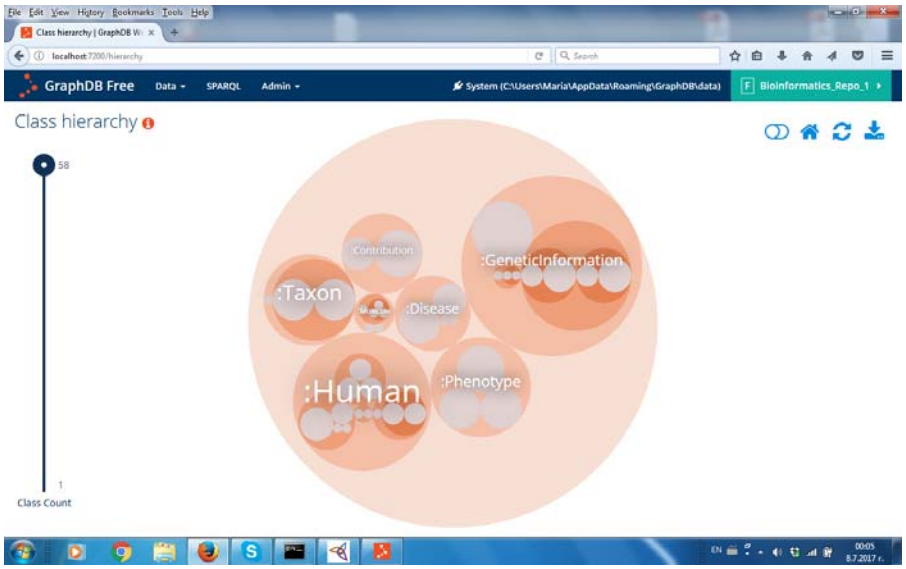

Fig. 1. Class hierarchy of BIO.

The core ontology describes hierarchies corresponding to five high-level concepts typically discussed in research and teaching resources in genetics (Molecule, Genetic Information, Genetic Contribution, Taxon, Phenotype) and two "purposive" concepts (Human and Disease) from genetics point of view. In the context of the basic concepts of genetics, the BIO ontology describes the human being (in the terms of the respective genes, allelic forms and chromosomes). On the basis of the available concrete data,

conclusions can be drawn by automated reasoners as to whether particular individuals have certain characteristics or are predisposed to specific diseases.

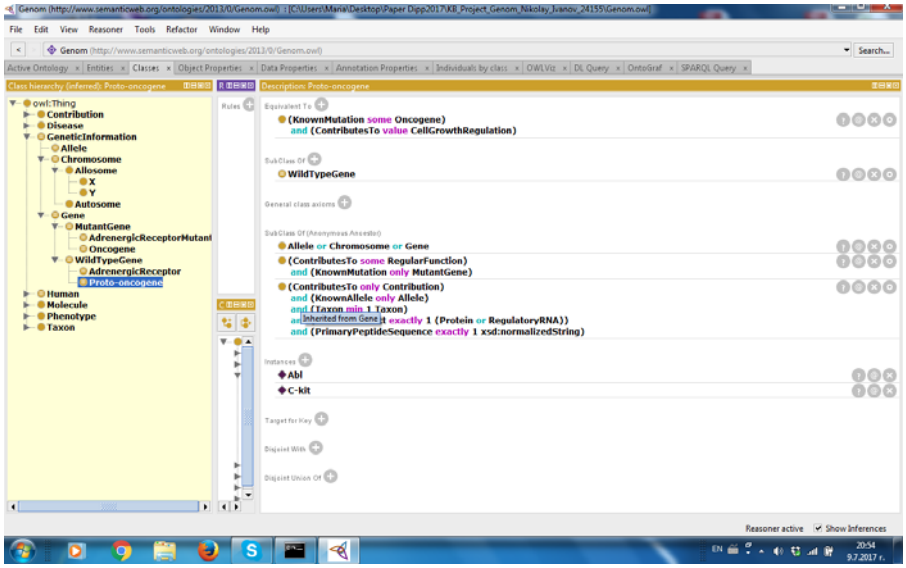

Fig. 2. Example class definition in BIO.

Most classes (concepts) of BIO are created as defined OWL 2 classes with necessary and sufficient conditions for belonging to them (see fig. 2). Multiple properties are described which define non-hierarchical relationships between classes (e.g. isPartOf, isAlleleOf, isMutationOf, hasChrmosome, hasGene, hasHeterozygousGene, etc.). These properties play a significant role in the semantic search process as a source of information for knowledge-based augmentation of user queries.

4 Conclusion and Future Work

Although our project is aimed at a particular domain, the final result of its implementation is expected to be sufficiently common and applicable to the creation of a broad class of semantic repositories with dynamically incoming heterogeneous resources. The accumulated experience will be a good basis for proper generalization and building a methodology for semantic enhancement of big collections of resources from interdisciplinary research.

5 Acknowledgements.

This work has been partially supported by the National Science Fund of Bulgaria within the "Methods for Data Analysis and Knowledge Discovery in Big Sequencing Datasets" project, contract I02/7/12.12.2014, and "Concepts and Models for Innovation Ecosystems of Digital Cultural Assets" project, contract DN02/06/15.12.2016.

References

Hoehndorf, R., Schofield, P., & Gkoutos, G. (2015). The Role of Ontologies in Biological and Biomedical Research: A Functional Perspective. *Briefings in Bioinformatics, 16*, 1-12.

Huang, J., et al. (2010). Ontology-Based Knowledge Discovery and Sharing in Bioinformatics and Medical Informatics: A Brief Survey. *Proceedings of Seventh International Conference on Fuzzy Systems and Knowledge Discovery FSKD 2010. 5*, pp. 2203–2208. IEEE.

Nisheva-Pavlova, M., & Pavlov, P. (2009). Building a Digital Library with Learning Materials. In S. Mornati, & T. Hedlund (Ed.), *Proceedings of the 13th International Conference on Electronic Publishing, "Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies"* (pp. 471-483). Milan, Italy: Edizioni Nuova Cultura – Roma.

Nisheva-Pavlova, M., Shukerov, D., & Pavlov, P. (2015). Design and Implementation of a Social Semantic Digital Library. *Information Services and Use, 35*(4), 273–284.

Schuurman, N., & Leszczynski, A. (2008). Ontologies for Bioinformatics. *Bioinformatics and Biology Insights, 2*, 187–200.

Stanchev, P., Nisheva-Pavlova, M., & Geske, J. (2010). Teaching Materials Repository. *Serdica Journal of Computing, 4*(3), 371–384.

Whetzel, P., et al. (2006). The MGED Ontology: A Resource for Semantics-based Description of Microarray Experiments. *Bioinformatics, 2*, 866–873.

Received: July 13, 2017
Reviewed: July 24, 2017
Final Accepted: August 02, 2017