Semantic search framework for distributed semantically based cheminformatics and bioinformatics datasets

Branko Arsić*, Marija Đokić*, Vladimir Cvjetković*, Petar Spalević**, Siniša Ilić**
* Faculty of Science, Kragujevac, Serbia
** Faculty of Technical Sciences, Kosovska Mitrovica, Serbia
brankoarsic@kg.ac.rs, m.djokic@kg.ac.rs, vladimir@kg.ac.rs, petar.spalevic@pr.ac.rs, sinisa.ilic@pr.ac.rs

***Abstract*— Integration of dissimilar and heterogeneous data sources is a continuous challenge for life science investigation. For the researchers in a domain of life science, finding the relevant information across different data sources is of crucial importance. Current researches strongly depend on availability, accessibility and effective data use. Semantic Web technologies enable facilitated collected data aggregating. Extracting data from large, semantically based datasets became a trying challenge for a researcher who needs to put an effort in discovering each connection in the datasets of interest, different properties among datasets and classes, name conventions, as well as to understand precise meaning of every enumerated part. For systems without specific conventions in organizing separate datasets, these tasks can be quite complex and time-consuming. In this paper, we present the challenges of data integration of same-domain datasets and the use cases that identify our approach aimed at finding the relevant data essential for our further research. We have developed a semantically based web application that utilizes the ontologies and SPARQL queries for the data process search and integration in molecular biology research center. For selected substance from our semantic dataset we have generated SPARQL queries for data discovering by means of the existing and constructed templates. In this way a knowledge database for valid and tested SPARQL queries is created and presented in a new ontology, established for this purpose.**

I. INTRODUCTION

Different institutions present their data in different manners, using diverse nomenclature and structure presentation. The data with equivalent meaning are presented in various formats and storages. A significant amount of knowledge is not available to all biomedical researchers, even though they have a frequent need to use datasets derived from other distributed systems. Most traditional integration methods are not scalable and based on static mapping approaches which results in significant access to information processing constraints. The most serious issue, caused by the constant expansion of novel data sources, is analyzing of disconnected data, which has become a threat for future successful and purposeful explorations. Such a quick-growing environment asks for updates monitoring worldwide. Bearing in mind all the things previously written, it is evident that the institutions with similar goals are faced with challenges when finding and comparing published data due to the fact that it is often necessary to interpret large amounts of data.

The current integration systems possess a lot of shortcomings such as inconsistent terminology, various data formats and customary alterations in data models. These shortcomings are transferred to the domain of biology and chemistry as well. Certain shortcomings in the domain of interest are dealt with in the section 3 with emphasis on difference between separated datasets. Semantic Web technologies have mechanisms to reduce the burden of data integration and sharing [1]. Semantic Web offers to its users well-defined models for data aggregation of heterogeneous data sources using explicit semantics among notions, with clear interlinks and relations - all of these packed with simplified annotation and with the possibility of publishing knowledge on web. With precise notion meaning, we receive an opportunity to connect data in different locations, unrelated at first glance, but sharing the biological and chemical domains. Ontologies are created as mechanisms for connecting similar data sources, offering a possibility to search heterogeneous datasets through a single SPARQL query [2][3]. Combination of ontologies and federated SPARQL queries for data retrieval can significantly simplify modeling of arbitrary concept and data structures and implementation of required functions. Numerous organizations use semantic web technologies to build ontologies as assistance in data integration and search processes. For instance, large initiatives (EBI, DrugBank, OpenPHACTS, PubChem,…) present their data by means of Semantic Web context, since they constantly updating the data structure.

Real system that is supported in this paper is the Research Center (RC) for testing of active substances [4]. The Center is also the forerunner of a large Project financed by the Ministry, and the subject of its analysis includes monitoring of in vitro effects of active substances in the cell lines of different origin (primarily cancer cell lines) and primary cells isolated from different tissues [5]. Active substances that are tested in laboratories are candidates for medicaments prior to being approved for medical treatments. Tests include measuring of the effectiveness of a substance in inhibiting a specific biological function (IC50) in human cancer cell lines, the mechanisms of apoptosis, migration and angiogenesis. The results of work in Center are experiments with complex structure that present complex relationships among various terms and concepts from the

Center work area. The structure is expected to further expand in the future, so it requires flexible modeling and representation that can be easily updated. These observations have led us to the semantic web technologies as appropriate choice. References [6][7] present earlier developed PIBAS (Preclinical Investigation of Bioactive Substances) ontology for data storing of active substances used in complex experiments, model systems, cancer cell-lines and experimental results.

Nowadays data sources are being developed in individual form isolated one from another. This leads to heterogeneous and challenging compute environment. Process of finding information about entity of interest in distributed systems can be difficult and laborious. Researchers are forced to follow the path between poorly connected sources. To achieve this aim, data about cell-lines, targets, proteins, pathways, drugs, and chemical compounds (substances) must be efficiently integrated and accessible to all the researchers. If a researcher wants to create valid SPARQL queries, he will have to discover almost all possible interlinks between notions. For many researches, this task can be a real challenge. Process of discovering new drugs, binding targets with cell lines, connecting pathways with active substances or compounds, relating genes with different diseases and similar processes can be very difficult.

Real need for application upgrading comes from the fact that some laboratories can test the same substance, but with different cancer-lines in different conditions. This complementary data can be very useful for QSAR analysis and very good direction for the future experiments. Some initiatives can be focused in other experiments and in other parameters such as pathways and targets. Obtained information is precious, because the search process has tendency to save researchers' time and resources. It will be shown later in examples how to get various experimental results for the same substance. In order to help life science community in finding relevant information we developed web application based on Semantic Web technologies. This application enables searching process within cheminformatics and bioinformatics datasets, based on federated SPARQL queries. Consequently, the application allows easy extraction of relevant information from all available, distributed sources. In the same time, this presents one kind of integration between datasets of interest. General endpoint is created, so in one place we have a possibility to explore some large systems in specific manner (see section 4). Everybody can create the templates and make them available for research community.

This paper is organized in the following way: The second section gives an overview of the literature and software in the domain of data integration in biology and chemistry. The third section talks about challenges in data integration and motivation for this work. The fourth section describes software's architecture as data integration mechanism between local ontologies and distributed systems that use templates and developed ontology. One part of this section is dedicated to federated SPARQL queries as a main product of software. Conclusion contains short survey of paper key points and directions for future work.

II. Related Work

Semantic Web technologies have potential to bridge before mentioned difficulties because they offer a common framework which permits data to be shared and reused across different systems. These technologies have a promising role in the field because they enable data integration and interlinking. Every initiative dealing with biology and chemical data has developed tools for data visualization and extracting on its own. In the following paragraph we present some of these large integrated systems and tools from our domain. Wild et al. indicated the importance of data integration in cheminformatics and bioinformatics [8]. Examples of successful exploitation and integration using Semantic Web technologies in biology can be seen in papers [9][10]. Ontologies as a main part of Semantic Web are finding their way in many areas of life sciences, especially in biomedicine [11].

Various initiatives for data integration of chemical and biological sources using a Semantic Web context have been established in the past decade. Open PHACTS represents Semantic web approach for addressing bottlenecks to drug discovery, developed as a shared platform for integration [12]. The European Bioinformatics Institute (EBI) provides freely available data from life science experiments, performs basic research in computational biology [13]. EBI developed Java based web application LODEStar as a generic SPARQL endpoint and Linked Data browser to provide a consistent interface and some enhanced functionality for querying and browsing EBI based dataset. Among many other services and tools, EBI offers UniChem API [14] as free available service which allows mappings of small molecule based on adopted and stable standard, InChIs and InChIKeys. Chem2Bio2RDF [15] is a solution based on Semantic web technologies which covers around 25 different datasets related to chemical/biology needs. This solution includes data about genes, compounds, drugs, pathways, side effects, diseases, and MEDLINE/PubMed documents. SLAP [16] tool for drug prediction is made by the same initiative. An aim of OpenTox community is to develop an interoperable predictive toxicology framework which may be used as an enabling platform for the creation of predictive toxicology applications [17]. For these purposes ToxPredict (http://www.toxpredict.org/) and ToxCreate (http://www.toxcreate.org/) applications are developed. Multiple chemical-protein annotation resources integrated with diseases and clinical outcomes information are presented by ChemProt integration system [18]. In recent years, various biological and chemical initiatives resulted in many tools and frameworks, including ChESS [19], WENDI [20], Data Dryad [21], IsaTab [22] and OpenTox.

Among large systems there are many identical datasets included. With this application we have direct access to all datasets which is positive since we can never be sure if a dataset is up to date.

III. Challenges in Integration of Biological and Chemical Data

Refined knowledge discovery process in life science includes moving through large numbers of interlinks among variant data sources. Researchers need to access several databases to perform tasks and identify potentially

constructive data, following each change manually, which is a time-consuming and error-prone process.

Current level of data integration is often facing syntactic and semantic heterogeneity challenges. The appearance of the same dataset in more than one global integration systems with different name convention, for the same compound, is probably the most common problem. For example, compound ZINC00120249 in Chem2Bio2RDF is denoted as 65-22-5 in ChemSpider integration model. With this notation we cannot be sure, without checking, if these substances are same or not.

In many situations compounds found in Chem2Bio2RDF and in our dataset are not available in some other dataset. This is caused by the fact that datasets within large system are not up to date or they have not treated substance yet. We need to be careful in creating queries because the results can be empty sets and thus get a wrong impression. Absence of substance in a result of a query is a good sign that tested substance is new on the market and that we are on the right track perhaps.

Problem in the forming of SPARQL queries is often followed by the absence of adequate endpoints. Many of them can be unsuitable, temporally unavailable and cause the lack of adequate results. In that case the problem is solved by using valid, alternative endpoints.

Large world initiatives take a few base datasets at the beginning and create the one-way relations to other internal datasets saving resources. Different integrated systems can have connected datasets but with different relation direction. In almost all situations one correct query is not correct for other systems. At the same time, different names and properties are used for the same notions.

IV. System Architecture

Our approach represents a collection of the templates from different initiatives from a domain of interest which are connected within developed core ontology. This approach enables an easier and faster way for searching included data sources than existing applications which are focused on a system which they belong to, rather than encompassing several initiatives in the same time. Every semantically based initiative invested a lot of time for presenting their work through the different SPARQL queries. Here, we want to avoid exploring datasets by using existing and self-created SPARQL queries. By the term "template" we shall mean every take over, valid SPARQL query. We gathered and connected them into unique storage in the form of ontology. New templates formed during the period of exploring these datasets are added as well.

Architecture of our web application (see Fig. 1) can be divided into three logical parts (layers): 1. User-interface for selecting input parameters and target integrated systems. Next step is filtering and selecting possible templates for selected systems 2. SPARQL framework represents an engine for adjusting templates into one federated query 3. Query execution on FUSEKI server and data retrieving. In the following paragraphs these layers are elaborated.

In Fig. 2 we can see the application interface. As possible input, molecular formula and weight are general parameters and the results offer the data of several substances. Sometimes this can cause confusion and make it difficult to track specific information. There are also two notations of chemical structures: *SMILES* and *InChI/InChIKey* strings. SMILES strings are omnipresent in chemical world, but they do not cope with every needed entity specification. Also, they are frequently not unique and cause the relevant data to misplace during the search. Conversely, the InChI and InChIKey (hashed InChI) strings lately gained a leading role among unique identifiers. Number and type of input parameters depend on the existing template's options and can be extended in the future.

As mentioned earlier, semantically based web application is developed with pillars in form of ontology (see Fig. 3). Developed ontology consists of information such as URI, endpoint and tested patterns, for every integrated system. Presented semantically based integrated systems have extendable list of possible templates for extracting ontology structures. For every template three parts pattern is constructed: a) *description* – used for user interface as a description of what we obtain as a result by embedding corresponding query b) allowed input parameter c) *query (template)* – SPARQL code with empty input parameter value. With input parameter choice and selected target system, web application filters possible corresponding templates that can be used in SPARQL generator layer. There is an option to choose several input parameters and integrated systems at the same time. Some of the integrated systems which are in focus in this paper are Chem2Bio2RDF, Bio2RDF, CHEBI, ChEMBL, DrugBank, LODD and

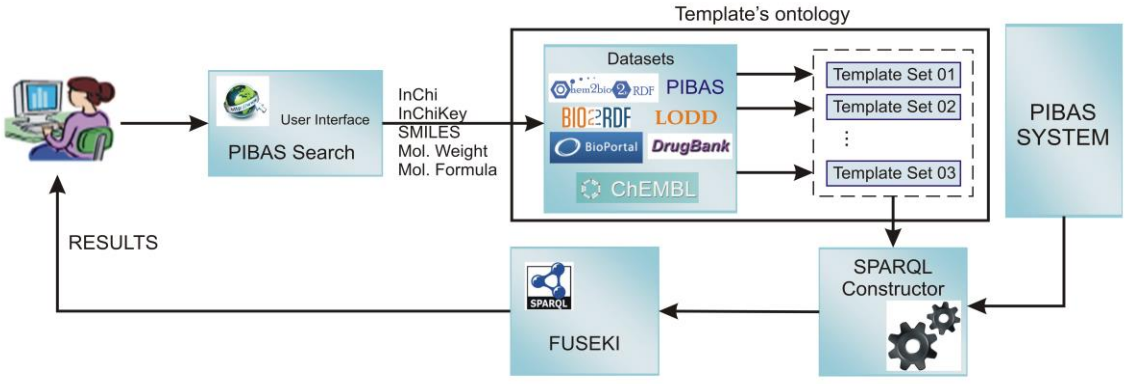

Figure 1. Web application arhitecture. Selected parameters determine final SPARQL query.

BioPortal. Possible templates are limited programmatically and a user can select a template according to given description.

Selected first layer arguments are forwarded to SPARQL constructor which generates federated SPARQL query with our input parameters' value. An example of such queries can be seen in Fig. 4. If we select more systems, we will obtain result with data from several datasets. Some of the data are complementary, and some are redundant. Different systems can have different standards for the same notion and we cannot say with certainty, in constructor runtime, that one notion is equal to another. After query is generated, an application sends it to FUSEKI server with general endpoint for execution.

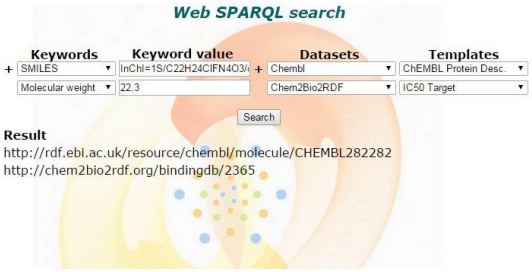

Figure 2. Application interface

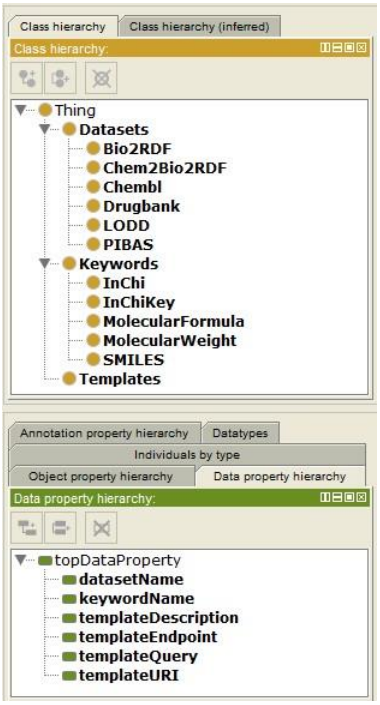

Figure 3. Ontology for integrated system's templates

V. CONCLUSION

The main activities of the Research Center include chemical synthesis, purification and extraction of bioactive substances, microbiological, cell and molecular, immunological and pharmacological preclinical testing of active substances. It has been proven, and in practice confirmed, that certain classes of chemical structures show larger or smaller biological effect. Data stated above can indicate whether the special attention should be paid to the new substance or not, thus avoiding time and resource consumption. The existence of such a database and of the corresponding software implies that the communication in the opposite direction exists as well, where CPCTAS could suggest to chemists a new synthesis direction. On one hand, in this way we can gain extra knowledge about substances through different kinds of experiments, in various conditions. On the other, we can have a new untested compound which could be important in terms of scientific research. The confirmation for this assumption stems from an empty result set. Fast accumulation of new, not-up-to-date data and reorganization of existing data can cause major problems in knowledge discovering. With new software we can facilitate the research work. Our application can be a standard for other small laboratories which want to improve the work and save resources.

For the future work we plan to deal with redundant data of the application results. For example, substance with molecular formula $C_{20}H_{12}N_2O_2$ and Chembl synonym 2'-Phenyl-[2,4']Bibenzooxazolyl is denoted as *CHEMBL113081*. The same substance within Chem2Bio2RDF initiative is denoted as *m180094*. However, this problem could be overcome with InChI/ InChIKey value. Different standards for the same notion are apparent in the case of data/object properties naming. Substance's target property within Bio2RDF initiative is *gene-name* from DrugBank namespace, and target property within Chem2Bio2RDF initiative is *geneSymbol* from Uniprot namespace. This problem is much bigger and requires deeper exploration of all the large systems and laborious mapping processes. On the other side, every initiative deals with different experiments focusing on specific results. Complementary data are also important, and with redundancy presents "all in one pack" problem with similar steps in their solving.

```
PREFIX pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#/>
PREFIX compound: <http://chem2bio2rdf.org/pubchem/resource/>
PREFIX pubchem: <http://chem2bio2rdf.org/pubchem/resource/>
PREFIX chembl: <http://chem2bio2rdf.org/chembl/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?compound ?cell_line_name
FROM <http://chem2bio2rdf.org/pubchem>
FROM <http://chem2bio2rdf.org/chembl>

WHERE{
?compound compound:std_inchi ?compound_inchi.
?compound_inchi pibas:InChi ?inchi.
?compound_inchi pibas:MolecularWeight ?mol_weight.

SERVICE<cheminfov.informatics.indiana.edu:8890/sparql>
{
?activities chembl:molregno ?compound;
            chembl:standard_value ?standard_value;
            chembl:standard_units ?standard_units;
            chembl:assay_id ?assay_id.

?assay2target chembl:assay_id ?assay_id.
?assay2target chembl:tid ?cell_line.

?cell_line chembl:pref_name ?cell_line_name.
}
FILTER regex(?cell_line_name,"cancer","i").
FILTER(?mol_weight="223.5").
FILTER(?inchi="InChI=1S/C22H24ClFN4O3/c1-29-20-13-19-16(12-21(20)31-8-2-5-28-6-9-30-10-7-28)22(26-14-25-19)
27-15-3-4-18(24)17(23)11-15/h3-4,11-14H,2,5-10H2,1H3,(H,25,26,27)").
}
```

Figure 4. Generated SPARQL query

ACKNOWLEDGMENT

This paper was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (scientific projects TR32023, III41010, ON174033 and III44006).

REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", Scientific American, vol. 284 (5), pp. 29-37, May 2001.

[2] N. Guarino, D. Oberle, and S. Staab, "What is an Ontology?", In *Handbook on Ontologies*, Springer Berlin Heidelberg, 2009.

[3] J. Pérez, M. Arenas, and C. Gutierrez, "Semantics and Complexity of SPARQL," In *The Semantic Web-ISWC 2006*, Springer Berlin Heidelberg, pp. 30-43, 2006.

[4] CPCTAS-LCMB, Faculty of Science, University of Kragujevac, Serbia, http://cpctas-lcmb.pmf.kg.ac.rs.

[5] Project data http://cpctas-lcmb.pmf.kg.ac.rs/lcmb/pibasEn.htm.

[6] V. Cvjetković, M. Đokić, B. Arsić, and M. Ćurčić, "The ontology supported intelligent system for experiment search in the scientific research center", *Kragujevac Journal of Science*, vol. 36, pp. 95-110, 2014.

[7] B. Arsić, M. Đokić, V. Cvjetković, P. Spalević, M. Živanović, and M. Mladenović, "Integration of bioactive substances data for preclinical testing with Cheminformatics and Bioinformatics resources," *Proceedings of the 23nd International Electrotechnical and Computer Science Conference (ERK 2014)*. pp. 146-149.

[8] D. J. Wild et al., "Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research," *Drug discovery today*, vol. 17(9), pp. 469-474, 2012.

[9] H. Min et al., "Integration of prostate cancer clinical data using an ontology," *Journal of biomedical informatics*, vol. 42(6), pp. 1035-1045, 2009.

[10] D. Salvi et al., "Merging Person-Specific Bio-Markers for Predicting Oral Cancer Recurrence Through an Ontology," *Biomedical Engineering*, IEEE Transactions on, vol. 60(1), pp. 216-220, 2013.

[11] B. Smith et al., "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology*, vol. *25*(11), pp. 1251-1255, 2007.

[12] A. J. Williams et al., "Open PHACTS: semantic interoperability for drug discovery," *Drug discovery today*, vol. 17(21), pp. 1188-1198, 2012.

[13] S. Jupp et al., "The EBI RDF platform: linked open data for the life sciences," *Bioinformatics*, vol. *30*(9), pp. 1338-1339, 2014.

[14] J. Chambers et al., "UniChem: a unified chemical structure cross-referencing and identifier tracking system," *Journal of Cheminformatics*, vol. 5(3), 2013.

[15] B. Chen et al., "Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data," BMC bioinformatics, vol. 11(1), 255, 2010.

[16] B. Chen, Y. Ding, and D. J. Wild, "Assessing drug target association using semantic linked data," *PLoS computational biology*, vol. 8(7), e1002574, 2012.

[17] B. Hardy et al., "Collaborative development of predictive toxicology applications," *J. Cheminform*, vol. 2(1), 2010.

[18] O. Taboureau et al., "ChemProt: a disease chemical biology database," *Nucleic acids research*, vol. 39, pp. 367-372, 2011.

[19] L. L. Chepelev and M. Dumontier, "Chemical entity semantic specification: knowledge representation for efficient semantic cheminformatics and facile data integration," *J. Cheminform.* vol. 3(20), 2011.

[20] Q. Zhu et al., "WENDI: a tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications," *J. Cheminform.* vol. 2(6), 2010.

[21] J. Greenberg, "Theoretical considerations of lifecycle modeling: an analysis of the dryad repository demonstrating automatic metadata propagation," *Inheritance, and value system adoption. Catalog. Classif. Quart.* vol. 47, pp. 380–402, 2009.

[22] S. A. Sansone et al., "Toward interoperable bioscience data," *Nat. Genet.* vol. 44, pp. 121–126, 2012.