Querying Phenotype-Genotype Associations across Multiple Knowledge Bases using Semantic Web Technologies

Oya Deniz Beyan, Aftab Iqbal, Yasar Khan, Athos Antoniades, John Keane, Panagiotis Hasapis, Christos Georgousopoulos, Myrto Ioannidi, Stefan Decker, Ratnesh Sahay

***Abstract*—Biomedical and genomic data are inherently heterogeneous and their recent proliferation over the Web has demanded innovative querying methods to help domain experts in their clinical and research studies. In this paper we present the use of Semantic Web technologies in querying diverse phenotype-genotype associations for supporting personalized medicine and potentially helping to discover new associations. Our initial results suggest that Semantic Web technologies has competitive advantages in extracting, consolidating and presenting phenotype-genotype associations that resides in various bioinformatics resources. The developed querying method could support researchers and medical professionals in discovering and utilizing information on published associations relating disease, treatment, adverse events and environmental factors to genetic markers from multiple repositories.**

I. Introduction

In this paper we present the use of Semantic Web technologies for querying diverse bioinformatics resources linking genotypes to phenotypes. Advances in genomics research have led to rising expectations for the arrival of the personalized medicine era expected to increase benefits and reduce risks for patients [1]. During the last decade, research studies in various domains such as bioinformatics, pharmacogenomics, and clinical research have focused on the personalization of health care from prediction and early detection of diseases to individualized treatment. Recent technologies have reduced the cost of high throughput genetic testing providing high density genetic data, however this data is hard to interpret and translate into meaningful information even for medical professionals [2]. Integrating large volumes of genomic information into medical practice is not straightforward. Harnessing knowledge from scientific research to produce new drugs and treatment options for patients as well as to enable early detection of adverse events requires innovative interoperability frameworks and business models to support information flow between scientists and research institutions [3,4].

In the work presented in this paper we study how Semantic Web technologies can be applied in combining phenotype genotype association across heterogeneous multiple knowledge bases for mining genetic associations. A primary aspect of Semantic Web technologies is a mechanism for defining and linking heterogeneous data using Web protocols and a flexible data model. In this work, the Resource Description Framework (RDF) is utilized. At the same time, feature selection and identification of the most predictive attributes is a key component for building successful models in bioinformatics and genomic medicine domains [5,6]. Semantic Web technologies enable use of data mining and knowledge extraction tasks in these domains by enabling researchers to effectively use prior domain knowledge through the querying of available knowledge bases.

The work presented in this article is developed as part of the Linked2Safety[1] EU project. Linked2Safety aims to facilitate clinical practice and accelerate medical research by providing researchers, pharmaceutical companies and healthcare professionals with a scalable technical infrastructure and a platform for effective utilization and reuse of semantically-interlinked medical information resources [7].

Within the scope of the Linked2Safety project, it is critical to provide tools to researchers for selecting the most significant attributes in their genetic data analysis. There are sets of scenarios including analyzing the associations between genotype data related to reported adverse events in clinical trials and identification of relations between molecular fragments and specific adverse event categories [8]. Utilizing prior domain knowledge on genotype phenotype associations should significantly leverage outcomes of those scenarios.

The aim of this research is to integrate distributed knowledge sources by employing Semantic Web technologies to support the needs of both researchers and clinicians at different levels. In order to demonstrate use of

Manuscript received July 30, 2013. This work is partly funded by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 and EU project Linked2Safety (contract Number 288328). Oya Deniz Beyan was supported by TUBITAK BIDEB 2219 grant during this study at NUI Galway DERI in Ireland.

O. D. Beyan is with the DERI National University of Ireland Galway, IRELAND (phone: +353 91 495053; fax: +353 91 495053; e-mail: oya.beyan@ deri.org).

A. Iqbal, Y. Khan, S. Decker, R. Sahay are with the DERI National University of Ireland Galway, Ireland (e-mail: {aftab.iqbal, yasar.khan, stefan.decker, ratnesh.sahay}@deri.org).

Athos Antoniades is with the University of Cyprus, Nicosia, Cyprus (e-mail: athos.antoniades@stremble.com).

J. Keane is with the University of Manchester, United Kingdom (e-mail: john.keane@manchester.ac.uk).

P. Hasapis, C.Georgousopoulos, are with the INTRASOFT International S.A., Luxembourg (e-mail: {Panagiotis.Hasapis, Christos.Georgousopoulos}@intrasoft-intl.com).

M. Ioannidi with the ZEINCRO HELLAS S.A., Athens, Greece(e-mail: mioannidi@zeincro.com).

[1] http://www.linked2safety-project.eu/

the developed framework we present two different use cases; (1) an examination of patient genetic data at the clinical level, (2) data mining at the research level.

II. Use Cases

A. Use Case 1: Risk Prediction and Disease Prognosis

The clinician required a personalized treatment plan for one of his patients diagnosed with Behcet Syndrome. He wanted to investigate any genomic cause associated with the current diagnosis that potentially might impact on the prognosis of disease and clinical outcomes. He had the genome sequence of patient as data; however there were numerous single-nucleotide polymorphism (SNP) variations which make difficult for him to examine and associate with current disease.

To deal with this situation, he had decided to conduct a literature review to identify possible variants related with Behcet Syndrome. He had access to the internet and utilized PubMed [9], to search for Behcet Syndrome-centered research publications hoping to find reported associations between genetic markers and disease outcome for specific treatment options. However, it is difficult to go through the mountain of information in all publications relating to Behcet syndrome and many of them are not open access. Further to that, he felt the challenge of translating the outcomes of these publications into his practice was too great for him to undertake especially since on some occasions contradictory or inconclusive outcomes were reported.

If he had been aware of genotype-phenotype association databases, he could have utilized them to quickly identify the genetic markers and the associations in which they are involved. He could have also identified the specific publications that led to these discoveries along with more recent ones that validated them.

After an extensive search, he identified a list of SNPs associated with Behcet Syndrome. However, the next challenge was to compare the patient's genomic data with the list of all susceptible variations. He successfully managed this task by using his computer literacy skills and finally identified his patient predisposition based on genetic markers rs1800871 and rs17810546.

The clinician also observed that the rs1800871 variation had been located at IL12A gene and rs17810546 variation is intergenic to RPS2P19 and IL12A. He wondered if there are any diseases that might be associated with these variations located on those gene regions. He conducted another literature review and discovered that there are some variations associated with Colitis, Ulcerative as well as Behcet Syndrome in IL12A gene region. Also there are some variations associated to Celiac Disease and Multiple Sclerosis in RPS2P19, IL12A intergenic region. Through his patient's existing genetic test that involved some of these markers, he was able to identify potential predisposition to other diseases. The physician felt the degree of certainty was too small to alarm his patient about his increased predisposition considering he had not diagnosed any phenotypes related to these diseases. However he utilized this information to make suggestions to his patient in order to reduce the environmental factors that contribute to these diseases as a way to reduce the probability of his patient acquiring them later on in life.

B. Use Case 2: Researcher – Exploring Genotype-Phenotype Associations over the Gene Pathways

Pathways are critical for elucidating mechanisms of diseases. Many complex diseases might target sets of genes involved in the same pathway. Researchers might be interested in exploring variations related to genes involved in the same pathways.

However this task requires significant effort in consolidating different knowledge sources and associating the derived information with the studied dataset.

Consider the case that a researcher had been investigating a dataset on diabetes mellitus type 1. She wanted to merge her data set with other related ones for data mining purposes. Her data set included a set of genomic data. Further she was looking for datasets that cover any genomic variation on the same pathway. Initially she focused on finding a list of related variations and their locations in genes for diabetes mellitus type 1 cases from genotype-phenotype association studies. Later, she had to review if any of those genes are included on any known pathways. Finally, she had to search related SNPs and diseases for each of the genes included in pathways. As the result of this labour-intensive literature review the researcher finally had a list of potentially related SNPs and diseases, which could guide her to find any other related datasets for data mining purposes.

III. Method

A. Genotype Phenotype Association Knowledge Bases

There are already numerous promising results obtained on complex analysis of genetic associations with disease characteristics and drug-target relationships. Genome wide association studies (GWAS) have identified many genetic loci that are associated with phenotypic traits and diseases. Various curated databases have started to publish genotype-phenotype associations studies results such as NCBI's dbGAP, GWAS Integrator, DistiLD database, GWASdb [10,11,12,13]. Biological systems database that integrates genomic, chemical and systemic functional information have been realized such as KEGG [14]. There are also other knowledge bases that curate knowledge about the effect on genetics on drug response such as PharmGKB [15].

In this work, we demonstrate the use of dbGAP, GWAS catalog data and KEGG knowledge bases.

B. Semantic Web Technologies

Despite the fact that those beneficial knowledge bases are publicly available, utilizing them in daily practice is still uncommon practice for clinicians and scientists. One of the reasons is the diversity of each knowledge base and the information contained within. Knowledge bases are widely distributed, maintained by various institutions or projects and represented differently to serve local needs. Retrieving and integrating information from those sources is time-consuming in daily practice. Hence, there is a need for a common model and standard format to represent a project's metadata from multiple knowledge bases to allow for better integration of results. One may think of questions like: what is the best way to express the knowledge so that it can be integrated easily across multiple knowledge bases? Can the knowledge be further used to link to other data sources which contain extra information about a certain entity? Can this be done in an automated fashion?

```
@prefix sehr: <http://hcls.deri.ie/l2s/sehr/genPheSEHR/1.0#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix so: <http://purl.org/obo/owl/SO#> .
@prefix acgt: <http://www.ifomis.org/acgt/1.0#> .
@prefix tmo: <http://www.w3.org/2001/sw/hcls/ns/transmed/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix snpo: <http://www.loria.fr/~coulet/ontology/snpontology/
              version1.6/snpontology_full.owl#> .

<http://www.linked2safety-project.eu/dbGap#40419>
   a      sehr:AssociationRule ;
   sehr:FunctionClass "Missense" ;
   sehr:KnowledgeBase "NHGRI GWAS Catalog" ;
   sehr:Position "67531642" ;
   sehr:PubMedID <http://www.linked2safety-project.eu/
                 pubmed#21829393> ;
   sehr:Pvalue
1.000000000000000062281591457779856418897068692785978 74E-9 ;
   so:SO-0000694 <http://www.linked2safety-project.eu/snp#rs763361> ;
   acgt:Chromosome "18" ;
   snpo:gene <http://www.linked2safety-project.eu/gene#CD226> ;
   tmo:TMO_0019 <http://www.linked2safety-project.eu/
                dbGap/trait#163> .

<http://www.linked2safety-project.eu/dbGap/trait#163>
   a      skos:Concept ;
   rdfs:label "Diabetes Mellitus, Type 1" .
```

Fig. 1. Representation of dbGap association rule in RDF format.

We propose the usage of Semantic Web technologies to represent data from different knowledge bases. As such, we propose to use RDF (Resource Description Framework) as the core, target data model. Further the RDF and OWL standards are used to define the vocabulary needed to describe the data (i.e., classes and properties). Once modeled in RDF, the data is indexed and queried using the SPARQL query standard and associated tools. Finally, the integrated data is published on the Web using Linked Data principles [16] allowing third parties to discover and subsequently crawl the knowledge, also allowing interlinking with background information available remotely on the Web. For space reasons, we refer the readers to [17] for details on how these standards can be used.

C. Transforming Knowledge Bases to RDF

In this section, we describe our methodology to extract information from different knowledge bases and the usage of a common model and standard format to represent extracted information in order to support better query and integration.

We downloaded KEGG [18] , dbGAP [19] and GWAS catalog [20] data from their respective sites and used our custom written scripts to convert them to RDF. An excerpt of an exemplary RDF representation of a dbGAP association is shown in Figure 1.

Due to space limitation we do not show the RDF representation of information extracted from other knowledge bases. After transforming the data sources to RDF, we loaded RDF datasets to our public SPARQL endpoint [21].

IV. Results

In this section we demonstrate and discuss how semantic technologies are applied to enable efficient investigation of genotype - phenotype associations within the scope of the use cases presented above.

In the first use case, the clinician requires associations between genotypes and phenotypes related to his patient's genomic profile. In this case our methodology would allow the clinician to run a single query on dbGAP and GWAS knowledge bases combining all the phenotypes related with his patient's genomic profile.

```
prefix tmo: <http://www.w3.org/2001/sw/hcls/ns/transmed/>
prefix so: <http://purl.org/obo/owl/SO#>
prefix sehr: <http://hcls.deri.ie/l2s/sehr/genPheSEHR/1.0#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix snpo: <http://www.loria.fr/~coulet/ontology/snpontology/
        version 1.6/snpontology_full.owl#>

SELECT ?s ?traitLabel ?rsLabel ?context ?geneLabel ?pvalue
?pubLabel {
   ?s a sehr:AssociationRule .
   ?s sehr:FunctionClass ?context .
   ?s snpo:gene ?gene .
   ?gene rdfs:label ?geneLabel .
   ?s sehr:PubMedID ?pub .
   ?pub rdfs:label ?pubLabel .
   ?s tmo:TMO_0019 ?trait .
   ?trait rdfs:label ?traitLabel .
   ?s so:SO-0000694 ?rs .
   ?s sehr:Pvalue ?pvalue .
{?rs rdfs:label "rs17810546" .} Union {?rs rdfs:label "rs1800871" .}
FILTER (?traitLabel = "Behcet Syndrome") .
}
```

Fig. 2. SPARQL query for retrieving association studies related with Behcet Syndrome and specific genetic profile

Patient genetic data is received in the variant call format (vcf) and the clinician additionally wants to limit his search to Behcet Syndrome. The system will receive those inputs and query the dbGAP and GWAS knowledge bases with the SPARQL query presented in Figure 2.

The following results display evidence returned to the clinician related to the patient's genomic profile:

Phenotype	Variation	Context	Gene	P values	PubMedID
Behcet Syndrome	rs17810546	Intergenic	RPS2P91	1.49E-05	20622878
Behcet Syndrome	rs17810546	Intergenic	IL12A	1.49E-05	20622878
Behcet Syndrome	Rs1800871	NearGene-5	IL10	1.00E-14	20622879

The clinician initiates a further query and gets other known variations associated with a disease in similar regions and a p-value of less than 10e-5. Figure 3 presents SPARQL query for IL12A – RPS2P19 gene region.

```
prefix tmo: <http://www.w3.org/2001/sw/hcls/ns/transmed/>
prefix so: <http://purl.org/obo/owl/SO#>
prefix sehr: <http://hcls.deri.ie/l2s/sehr/genPheSEHR/1.0#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix acgt: <http://www.ifomis.org/acgt/1.0#>
prefix snpo: <http://www.loria.fr/~coulet/ontology/snpontology/
             version 1.6/snpontology_full.owl#>

SELECT ?s ?traitLabel ?rsLabel ?context ?pvalue ?pubLabel {
    ?s a sehr:AssociationRule .
    ?s sehr:FunctionClass ?context .
    ?s snpo:gene <http://www.linked2safety-project.eu/gene#IL12A> .
    ?s snpo:gene <http://www.linked2safety-project.eu/gene#RPS2P19> .
    ?s sehr:PubMedID ?pub .
    ?pub rdfs:label ?pubLabel .
    ?s tmo:TMO_0019 ?trait .
    ?trait rdfs:label ?traitLabel .
    ?s so:SO-0000694 ?rs .
    ?rs rdfs:label ?rsLabel .
    ?s sehr:Pvalue ?pvalue .
    FILTER (?pvalue < 0.00001)
    FILTER (?context = "Intergenic")
}
```

Fig. 3. SPARQL query for retrieving association studies related with IL12A – RPS2P19 gene region

This query returns all the phenotypes associated to the variants in the intergenic region of IL12A- RPS2P19 above the specified evidence level from dbGAP as follows:

Phenotype	Variation	Context	P values	PubMedID
Celiac Disease	rs17810546	Intergenic	1.00E-09	18311140
Celiac Disease	rs17810546	Intergenic	4.00E-28	20190752
Multiple Sclerosis	rs4680534	Intergenic	6.00E-06	19525953

Query is repeated for all gene and intergenic regions. Variants associated with Colitis Ulcerative, Celiac Disease and Multiple Sclerosis, as well as Behcet Syndrome are retrieved.

Application of semantic technologies will facilitate the knowledge discovery process and the clinician would be able to focus on patient care rather than performing literature reviews.

In the second use case the researcher prepares a dataset for data mining. She needs to extract the genetic markers associated with adverse events for diabetes mellitus type 1. She wants to include variants in related pathways.

In this use case SPARQL queries interact both with the genotype – phenotype association knowledge base and the KEGG gene-pathway knowledge base to retrieve up-to-date results. In this use case we present how to retrieve genes and gene regions in a pathway and query each location returned from the phenotype-genotype association knowledge bases. Figure 4 presents the SPARQL query that returns gene in the hsa04940 pathway for Type I diabetes mellitus.

```
prefix sehr: <http://hcls.deri.ie/l2s/sehr/genPheSEHR/1.0>
prefix tmo: <http://www.w3.org/2001/sw/hcls/ns/transmed/>
prefix snpo:
<http://www.loria.fr/~coulet/ontology/snpontology/version
1.6/snpontology_full.owl#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?path ?trait ?gene {
    ?disease a sehr:KeggEntry.
    ?disease tmo:TMO_0019 ?trait.
    ?trait rdfs:label "Type I diabetes mellitus".
    ?disease tmo:TMO_0044 ?path.
    ?path rdfs:label "hsa04940".
    ?disease snpo:gene ?gene.
}
```

Fig. 4. SPARQL query returns gene in the hsa04940 pathway for Type I diabetes mellitus

The results of a query executed over the KEGG dataset returned 21 genes including {HLA-DRB1;HLA-DQB1; HLA-DQA1; INS; CTLA-4; PTPN22; IL-2RA; PTPN2; ERBB3; IL2 - IL21; IFIH1; CLEC16A; BACH2; PRKCQ; CTSH; C1QTNF6; SH2B3; C12orf30; CD226; ITPR3; CYP27B1}.

The returned list of genes is then used as a restriction condition for querying phenotype and genotype association knowledge bases. If the p value restriction is not provided, the number of reported variations is 45 for 32 different phenotypes from 91 association studies. However, when the researcher applied a p-value restriction smaller than 10e-15 and re-executes the query it returns 11 related variations for 10 different phenotypes from 29 association studies.

V. Related Work

In bioinformatics and biomedical domains there are many publicly available datasets that provide results of research studies at various levels including genes, human variations, proteins, proteomics, pathways and pharmacogenomics. Semantic Web technologies provide opportunities to develop tools for mapping and merging diverse heterogeneous data sources. Recently there have been several studies which employ some of those technologies towards achieving different goals. Pathak et. al. employs linked data to integrate Online Mendelian Inheritance in Man (OMIM) and Database for SNPs (dbSNP) within the LOD knowledge base for searching disease-gene and disease-SNP

associations for chronic diseases [22]. In another study, they used RDF to represent health records and genotype data, and then implemented a SPARQL-based querying interface [23]. In other approaches the Semantic Web Rule Language (SWRL) and Semantic Query Web Rule Language (SQWRL) are employed to query relevant phenotype-genotype bidirectional relationships over patient datasets and OMIM [24]. Another example of usage of semantic technologies is the development of a semantic model for pharmacogenomics information. In this study the designed model is mapped using Protein Ontology, HUGO Gene Nomenclature Committee (HGNC) and PharmGKB [25].

Our approach is also based on the Semantic Web principles and provides additional features. Firstly, we have extended the Semantic EHR ontology [8] developed within the scope of the Linked2Safety project to cover phenotype-genotype associations. As our semantic model is an ontology, it has the advantage of the representation of data from various knowledge sources in a consistent manner. Secondly, we have represented data using RDF and employed queries over SPARQL end points. With this approach we can extract and integrate data from multiple knowledge sources. Moreover our approach provides flexibility to integrate data at different levels with heterogeneous formats. Currently we have covered dbGAP, GWAS and Kegg pathway databases, and this work can be easily extended to other knowledge sources. This is a key advantage as research in genetics is already focusing on next generation sequencing technologies (NGS) that have the potential to completely sequence the entire genome of subjects involved in epidemiological studies. This should enable identification of not just SNPs (the primary focus of current databases) but also other genetic markers associated to diseases. Examples include CNV regions, insertion or deletions of bases (indel) etc. The approach proposed in this paper has been designed in such a way to enable easy incorporation of current and future data sources that may link diverse types of genetic or epigenetic markers to disease predisposition, treatment options, prevention of adverse events etc.

VI. CONCLUSION

In this work we have presented two different use cases that focused on the needs of clinicians and researchers and integrated three different knowledge bases namely dbGAP, GWAS and KEGG. In future, we plan to extend our work in two different directions. In order to support clinical usage we will integrate new data sources of clinical importance [26]. At the same time we will expand the number of integrated databases to include ones that will contain different types of genetic markers such as indels and CNVs.

REFERENCES

[1] M.J. Khoury, M Gwinn, P.W. Yoon, N Dowling, C.A. Moore, L. Bradley, “The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?” Genetics in Medicine, vol. 9, pp. 665–674., 2007 .

[2] U.O. Njiaju, OI.Olopade, “Genetic determinants of breast cancer risk: a review of current literature and issues pertaining to clinical application”. *Breast J*. 18(5),pp.436-42,|Sept.2012.

[3] H. Steven, M.D. Woolf, “The Meaning of Translational Research and Why It Matters”, JAMA, 299(2), pp.211-213, 2008.

[4] A. Antoniades, C. Georgousopoulos, N. Forgo, A. Aristodimou, F. Tozzi, P.Hasapis, K. Perakis, T. Bouras, D. Alexandrou, E. Kamateri, E. Panopoulou, K. Tarabanis, and C. Pattichis, “Linked2Safety: A secure linked data medical information space for semantically-interconnecting EHRs advancing patients' safety in medical research," in Proc. 2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE), 2012, pp. 517 -522.

[5] Y. Saeys, I. Inza, P. Larran˜aga, “A review of feature selection techniques in bioinformatics”, Oxford Journals Life Sciences & Mathematics & Physical Sciences Bioinformatics, Vol. 23, Issue 19, pp. 2507-2517, 2007.

[6] R. Bellazzi, B. Zupan, “Predictive data mining in clinical medicine: Current issues and guidelines”, INT J MED INFORM, vol 7, 7, pp. 81–97, 2 0 0 8.

[7] A. Elias, M.D. Zerhouni, “Translational and Clinical Science — Time for a New Vision”, N Engl J Med, 35, pp.1621-1623, 2005.

[8] R. Sahay, D. Ntalaperas, E. Kamateri, P. Hasapis, O. D. Beyan, M. F. Strippoli, C. A. Demetriou, T. Gklarou-Stavropoulou, M. Brochhausen, T. Tarabanis, T. Bouras, D. Tian, A. Aristodimou, A. Antoniades, G. Georgousopoulos, M. Hauswirth, and S. Decker "An Ontology for Clinical Trial Data Integration," In IEEE SMC 2013 - IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, October 13-16. IEEE Xplore, 2013

[9] NCBI PubMed, Available: http://www.ncbi.nlm.nih.gov/pubmed

[10] NCBI dbGAP, Available: http://www.ncbi.nlm.nih.gov/gap

[11] GWAS Integrator, Available: http://www.hugenavigator.net/HuGENavigator/gWAHitStartPage.do

[12] DistiLD database, Available: http://distild.jensenlab.org/

[13] GWASdb, Available: http://jjwanglab.org:8080/gwasdb/

[14] KEGG for linking genomes to life and the environment. Nucl. Acids Res. (2008) 36 (suppl 1): D480-D484.doi: 10.1093/nar/gkm882

[15] PharmGKB: the Pharmacogenetics Knowledge Base. Nucl. Acids Res. (2002) 30 (1):163-165.doi: 10.1093/nar/30.1.163

[16] Linked Data, Available: http://www.w3.org/DesignIssues/LinkedData.html

[17] T. Heath and C. Bizer, “Linked Data: Evolving the Web into a Global Data Space (1st edition)”, Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1):1-136, 2011.

[18] Kegg pathway: ftp://ftp.genome.jp/pub/kegg/medicus/disease/disease

[19] NCBI dbGAP Association Result Browser: http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm

[20] National Human Genome Research Institute, Catalog of Published Genome Wide Association Studies:http://www.genome.gov/gwastudies/

[21] SPARQL end point, Available: http://linked2safety.hcls.deri.org:3030

[22] J. Pathak, R. Kiefer, R. Freimuth, C. Chute, “Validation and discovery of genotype-phenotype associations in chronic diseases using linked data”. Stud Health Technol Inform.180: pp.549-53, 2012.

[23] J. Pathak, R. Kiefer, R. Freimuth, S. J. Bielinski, C. Chute,” Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank”. Journal of Biomedical Semantics, 3:10, 2012.

[24] M. Taboada1, D. Martínez, B. Pilo, A. Jiménez-Escrig, P. N. Robinson, M. J. Sobrido.” Querying phenotype-genotype relationships on patient datasets using semantic web technology: the example of cerebrotendinous xanthomatosis” BMC Medical Informatics and Decision Making, 12:78, 2012

[25] Boyce, RD., Freimuth, RR., Romagnoli, KM., Pummer, T., Hochheiser, H., Empey, PE. Toward semantic modeling of pharmacogenomic knowledge for clinical and translational decision support. Proceedings of the 2013 AMIA Summit on Translational Bioinformatics. San Francisco, March 2013.

[26] NCNI:ClinVar, Available: http://www.ncbi.nlm.nih.gov/clinvar/