# BiomRKRS: A Biomarker Retrieval and Knowledge Reasoning System

**BAHADORREZA OFOGHI [1,2], GUILLERMO HUGO LOPEZ CAMPOS [2], KARIN VERSPOOR [1,3], FERNANDO JOSE MARTIN SANCHEZ [2]**

[1] National ICT Australia, Victoria Research Lab
[2] Health & Biomedical Informatics Centre
[3] Department of Computing and Information Systems

The University of Melbourne
Victoria 3010, Australia

{bahadorreza.ofoghi,karin.verspoor}@nicta.com.au
{guillermo.lopez,fjms}@unimelb.edu.au

## Abstract

The need for a system to effectively manage and retrieve biomarker information has become apparent to medical and biomedical scientists, as evidenced by the recent development of a number of biomarker information systems. To improve the functionality of such systems, we have developed a new biomarker information system will be discussed in this paper, a system that we refer to as *BiomRKRS*: A Biomarker Retrieval and Knowledge Reasoning System. In this paper, we introduce the general structure and characteristics of BiomRKRS. We will demonstrate how BiomRKRS employs existing ontologies in the biomedical domain to create a core integrated ontology for biomarkers as a standard vocabulary set for data storage and retrieval. When fully implemented, BiomRKRS will have functionality and utility that will far exceed that of related existing systems due to the incorporation of a knowledge reasoning system that will make logical and useful inferences in the process of semantically processing end-user queries.

*Keywords*: Biomarker, Ontology, Data Retrieval

## 1 Introduction

Biomarkers have become central to the current practice of medicine and are an active focus for biomedical and translational research (Olson, Robinson, & Giffin, 2009). The term biomarker (biological marker) was first introduced as a Medical Subject Heading (MeSH) term in 1989 as *measurable and quantifiable biological parameters (e.g., specific enzyme concentration, specific hormone concentration, specific gene phenotype distribution in a population, presence of biological substances) which serve as indices for health- and physiology-related assessments, such as disease risk, psychiatric disorders, environmental exposure and its effects, disease diagnosis, metabolic processes, substance abuse, pregnancy, cell line development, epidemiologic studies, etc.* (Vasan, 2006). The US National Institutes of Health defines a biomarker as *a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention* (BDW, 2001). We specifically adopt the latter definition for its brevity and conciseness. Focusing on Alzheimer's Disease and Multiple Sclerosis, Younesi et al. (2012) proposed that biomarkers may represent molecular, physiological, or structural features and therefore, can be in the form of genes, proteins, DNA, RNA, genetic changes (e.g., SNPs), blood cholesterol levels, or patterns of brain abnormality. Going beyond this definition, our focus is on biomarkers generally in molecular entities including proteins, DNA, RNA, metabolites, and all of the subclassifications of these categories.

Biomarkers have been used for diagnosis, treatment, prognosis, and staging of different categories of diseases, examples of which include the biomarkers for management of postmenopausal osteoporosis (Szulc & Delmas, 2008), prediction and monitoring of osteoporosis (Vasikaran et al., 2011), diagnosis and prognosis of rheumatoid arthritis (Carrasco & Barton, 2010), prognosis and prediction of breast cancer (Weigel & Dowsett, 2010), and treatment of cardiovascular disease (Vasan, 2006), just to name a few.

In addition to directly disease-related procedures, another particularly valuable use of biomarkers is in bridging the gap between the preclinical and clinical development of drugs and vaccines (Olson et al., 2009). Biomarkers can play a role in toxicity/adverse reaction prediction and the analysis of the therapeutic effectiveness of drugs, e.g., dose-response relationship analysis.

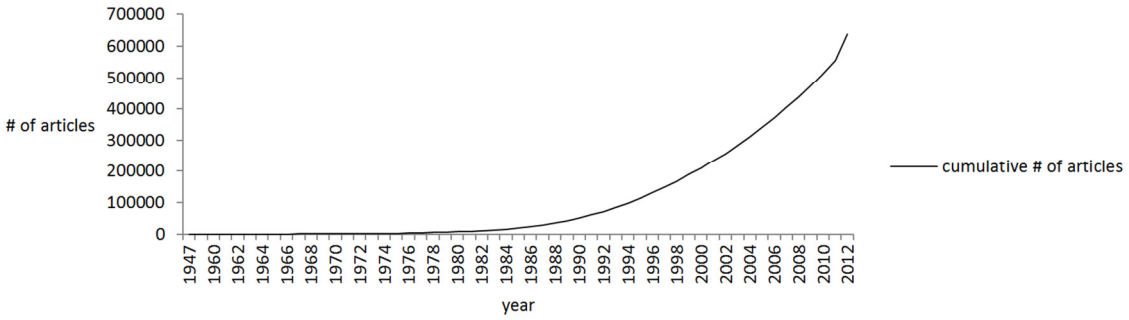


**Figure 1.** The cumulative number of articles published in PubMed in relation to biomarkers since 1947. The PubMed query is: "biological markers"[MeSH Terms] OR ("biological"[All Fields] AND ("markers"[All Fields] OR "marker"[All Fields])) OR "biological markers"[All Fields] OR "biological marker"[All Fields] OR "biomarker"[All Fields] OR "biomarkers"[All Fields] AND ("0001/01/01"[PDAT] : "2012/12/31"[PDAT])

In recent years, the U.S. Department of Energy Human Genome Project[2] and advances in genomic sequencing have enabled the detection of new biological features and entities that have been proposed as potential biomarkers for disease diagnosis, treatment, prognosis, and staging as well as for drug development. As a result, the size of the information being generated is increasing every year. To briefly demonstrate this growth of the amount of information related to biomarkers, we have extracted the total number of biomarker-related articles in PubMed from 1947 to 2012 using a simple keyword query. Figure 1 shows the result of this analysis in which the cumulative number of biomarker-related articles totals 638,885 for the specified time period. The diagram also shows how the number of these articles has substantially increased in recent years.

The growth in the number of studies addressing biomarkers necessitates the existence of highly efficient and effective information systems that enable fast and accurate search of and access to up-to-date biomarker information. Our definition of a biomarker information system is *a system that stores actual instances of biomarker information/data records as related to different contextual information attributes, e.g., the disease, clinical purpose, and molecular entity, just to name a few*. Such a system differs from databases of patient-related clinical information records in that the proposed biomarker information system does not include any patient-specific data but rather captures the background knowledge that relates to known biomarkers. This information can be applied to interpret patient-specific data, but is itself at a higher level of abstraction.

Some researchers have already started to develop biomarker databases and/or information systems for specific (categories of) diseases and drugs, including the following: a commercial collection of clinical, pre-clinical, and exploratory biomarkers named GVK BIO Online Biomarker Database (GOBIOM) (GVK Biosciences, 2013), the collection of validated molecular biomarkers in BiomarkerBase (Amplion Research, 2013), the set of population specific and clinically important biomarkers in Biomarker Databases (Liatris Biosciences LLP, 2013), the collection of biomarkers combined with related drugs, targets and genes in the evolvus Biomarkers Database (Evolvus, 2013), the standardized terminology and classification of biomarkers into lifecycle phases and disciplines in the biomarkers module of Thomson Reuters Integrity (Thomson Reuters, 2013), the set of diagnostic and prognostic cancer biomarkers extracted from patents, research articles, and meeting abstracts in the SciClips' Cancer Biomarker Database (SciClips, 2013), the knowledge-based interface for biomarkers for diagnosis, detection, protection, and characterization of infectious diseases developed in the Infectious Disease Biomarker Database (Yang et al., 2008), the services architecture for the capture, processing, management, and distribution of information in biomarker discovery and validation developed by Crichton et al. (2006), and the ontology representing concepts related to imaging biomarkers developed as Quantitative Imaging Biomarker Ontology (Buckler et al., 2013).

We found that the two major shortcomings of all the above information systems or knowledge-bases for biomarker data management are the following:

- The existing information systems make limited use of standard controlled vocabularies and ontologies for a comprehensive set of features that relate to biomarkers.
- None of the existing information systems makes use of logical reasoning functionalities available in the semantic web domain to semantically process user queries and retrieve (related) information for biomarker-focused user queries.

The advantage of adopting standard terminology and ontologies is two fold: i) the end-users (i.e., clinicians and biomedical researchers) will be able to interact with the system in a controlled manner where the concepts are well-established and well-known to the users, and ii) the system itself will have the ability to make use of reasoning engines over the well-formed structures of existing knowledge-bases. This in turn will enable the system to make inferences and retrieve not only exact matches to user queries but also logically related biomarker data. The latter can play an important role in identification of new indications to previously unknown biomarkers under certain conditions.

[2] http://www.ornl.gov/hgmis

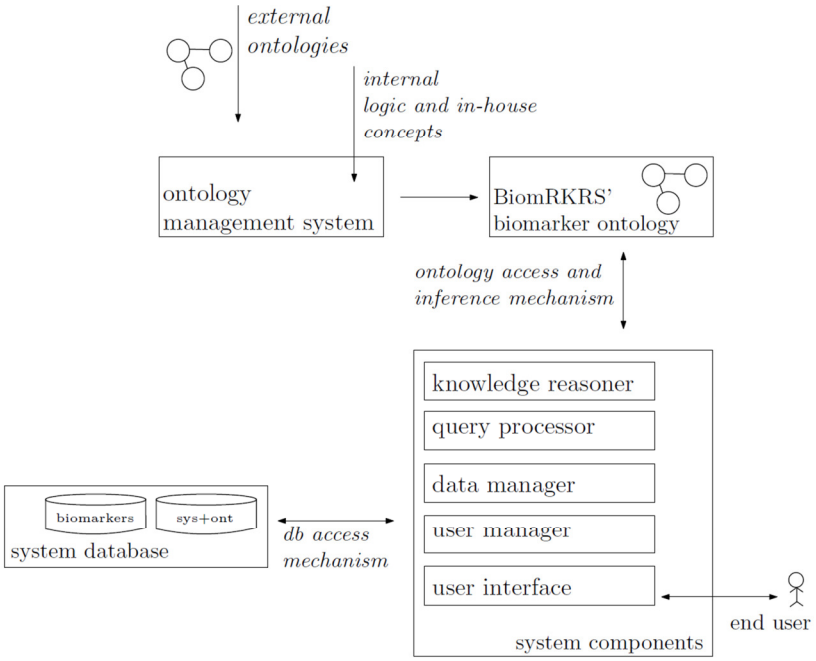


**Figure 2.** General architecture of BiomRKRS

In developing BiomRKRS (pronounced similar to *biomarkers*), consideration has been given to those two main drawbacks of existing biomarker information systems. In the next sections, the general structure of BiomRKRS will be introduced first and then, the main modules of the system will be discussed in more detail.

## 2 BiomRKRS: architecture

In order to address the two main shortcomings of existing biomarker information systems, BiomRKRS has been developed with a general architecture as shown in Figure 2. The system defines a core integrated biomarker ontology and expanded each core concept through reuse of a number of external ontologies. The core concepts and the external resources that relate to them are described in the next section.

There are some other internal concepts and related logic that are also created in order to construct the core ontology. This core ontology is then used mainly as: i) a (controlled) vocabulary resource for data storage and retrieval, and ii) a knowledge resource for inference purposes in BiomRKRS to semantically process end user queries.

The system implements several functionalities including a knowledge reasoner, a query processor, a data manager, a user manager, and a user interface. These components interact with each other as well as with the core ontology and the system database to answer information requests submitted by the end user. The system database stores three basic types of data: *semantic data* related to the core ontology, *transactional data* regarding system users and history, *biomarker instance data*, i.e., specific information about individual biomarkers.

In the following sections, the main components of BiomRKRS will be discussed in more detail.

### 2.1 The BiomRKRS biomarker ontology

Based on our expert knowledge, a number of concepts have been defined as related to the main and focal biomarker concept in the core ontology we have built for BiomRKRS. For each of these concepts, a specific external resource, i.e., an ontology in most cases, has been identified to provide specific terminology for the concept. Figure 3 illustrates the core ontology in BiomRKRS including the main concepts, their relationships with the focal biomarker concept, and the external ontologies and knowledge resources utilized for each core concept. In most cases, the external ontology or knowledge resource has been imported into the core ontology and then necessary relations have been created using OWL/XML statements between the external resource/ontology and the specific concept in the core ontology (indicated with an *imp* in Figure 3 on the links between an external resource and a main concept). In some other cases, the external ontology or resource has been used only as a reference and the actual concepts have been internally created in our ontology (indicated with a *ref* in Figure 3). More details on how the BiomRKRS ontology is constructed will be given in the next section. At this stage, the main concepts of the BiomRKRS core ontology include:

**disease:** the disease for which the biomarker is used for diagnosis or prognosis. For this concept, the International Classification of Diseases ICD-10 (World Health Organization, 1992) is used as the standard vocabulary.

**endpoint:** indicates whether the biomarker is a clinical or surrogate endpoint.

**molecular entity:** the main entity that is clinically measured as the main biomarker. For this concept, the lists of recommendations for molecular entities from the HGVS (Human Genome Variation Society, 2013) and HGNC terminology from HUGO Gene Nomenclature

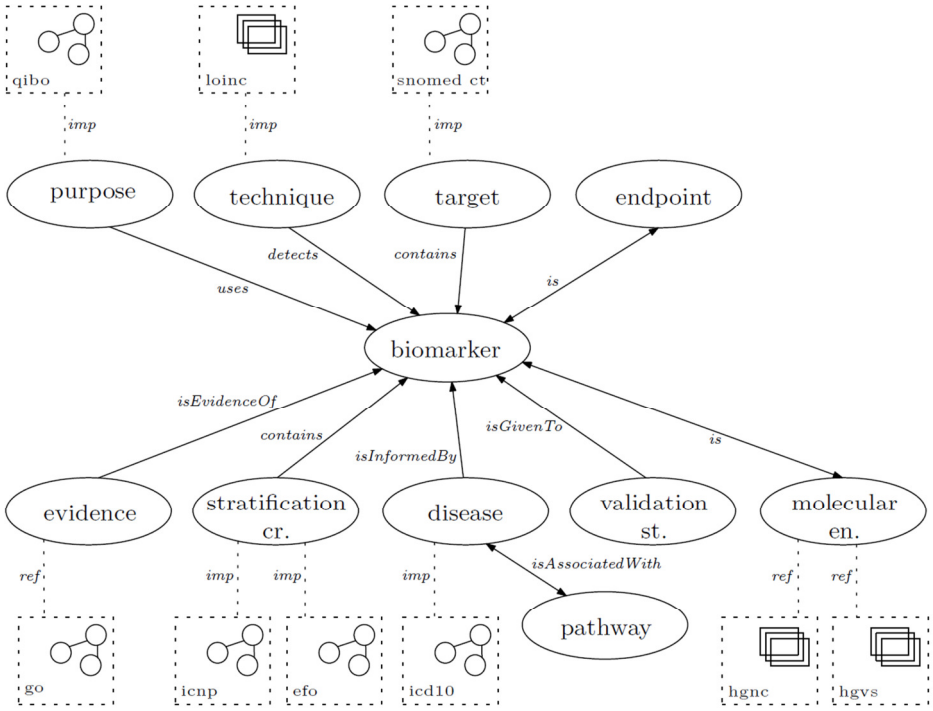


**Figure 3.** The BiomRKRS core biomarker ontology structure and related external resources. An *imp* indicates that the resource has been imported in the ontology whereas a *ref* shows that the resource has been used as a reference only. LOINC, HGNC, and HGVS are not ontologies and therefore, have been shown with different pictorial symbols.

Committee at the European Bioinformatics Institute (Gray et al., 2013) are used as the reference set, especially for the DNA and RNA entities.

**pathway:** the biological and genetic mechanism related to the specific disease for which the biomarker is measured.

**purpose:** the clinical purpose of measuring the specific biomarker. For a list of existing purposes, the Quantitative Imaging Biomarker Ontology (Buckler et al., 2013) is used. This ontology includes diagnosis, disease staging, and prognosis as purposes, *inter alia*.

**target:** the sample from the patient to be used in the clinical trial to measure the biomarker. For this concept, SNOMED Clinical Terms (Cornet & de Deizer, 2008) under the "specimen" term are used as the vocabulary.

**technique:** which represents the clinical technique used for measuring the specific biomarker in the specific target. The "method types" from Logical Observation Identifiers Names and Codes (LOINC) (Forrey et al., 1996) are used to expand this BiomRKRS concept.

**validation status:** the stage of validation and qualification of a biomarker. This attribute can have a value of *biologically validated*, *clinically validated*, *in research*, *proposed*, and *qualified* stages.

**stratification criteria:** characteristics that that could affect the validity or measure of the specific biomarker in patients. This includes the age group, race, and gender to which the specific biomarker is related as well as the environmental exposure factors. Concepts under the population term from the Experimental Factor Ontology (Malone et al., 2010) are used for the *race* concept in BiomRKRS. For environmental exposure, related concepts from the International Classification for Nursing Practice Ontology (International Council Of Nurses, 2013) are used.

**evidence:** the source of the biomarker information, e.g. related literature that suggests whether the specific biomarker is, with the specified validation status, to be considered as a measure or not to be used as a measure in the given context for the specific disease. There are also evidence codes defined in BiomRKRS similar to those from the Gene Ontology (The Gene Ontology Consortium, 2000), including computational evidence, experimental evidence, and their subclasses.

## 2.2 Ontology management system

Creating the BiomRKRS core biomarker ontology is the main task carried out by the ontology management system. This involves three main steps:

- Defining and constructing the core ontology, its main concepts, and necessary relationships, discussed in previous section.
- Defining and constructing the internal sub-ontologies for each core concept based on expert knowledge or reference resources.
- Importing external ontologies or knowledge resources and relating them to specific core concepts.

All the above tasks are performed in an automated function to make updates possible as new versions of external resources become available. End users will be able initiate updates through the user interface component, with appropriate access levels and permissions.

The core ontology and the main concepts are created using Web Ontology Language (OWL) in the OWL/XML format. For integration purposes, the external resources

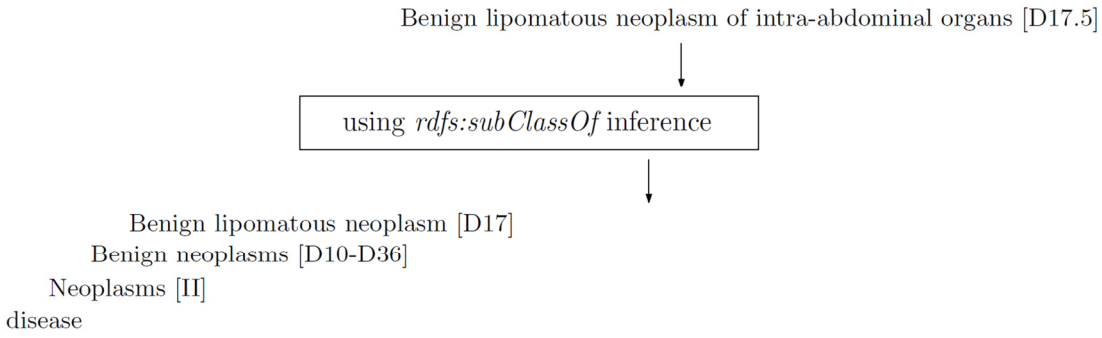


**Figure 4.** An example query term expansion using RDFS inferences

have also been imported in the same format. In cases where an OWL format was not readily available, we developed a converter functionality in (or prior to) the ontology management system.

In particular, we developed a lightweight converter function for the ICD-10 database to convert the database from its original format, i.e., Classification Markup Language (ClaML), into OWL. We made use of instructions given in (Moller, Sonntag, & Ernst, 2010) for most of the conversion procedure. We also developed a converter function to convert LOINC's Tab Delimited file into a Resource Description Format file as part of the ontology management system.

## 2.3 System components

As shown in Figure 2, there are five internal functionalities in BiomRKRS, implemented in separate components. These components handle accessing the BiomRKRS biomarker ontology, the system database, and interacting with end users.

**Knowledge reasoner** implements Resource Description Framework Schema (RDFS) reasoning capabilities. For this, the knowledge reasoner has access to the BiomRKRS ontology and its core and integrated concepts.

**Query processor** is the component that takes in original user query keywords and generates an expanded query. The expanded query includes the original terms of the user query as well as all keywords that are found related to the query term by using the knowledge reasoner component. Figure 4 shows an example query term expansion using the rdfs:subClassOf inference mechanism in RDFS. The selected disease category from the ICD-10 terms, i.e., *Benign lipomatous neoplasm of intra-abdominal organs [D17.5]*, has been found to be related to the three upper ICD-10 disease categories as well as the parent core concept "disease" from the BiomRKRS core ontology. The query processor component creates the expanded query using the "OR" operator between all the disease categories.

**Data manager** is the component that has been implemented to interact with the user interface and the system database. The data that this component handles are related to the system's history, ontology files and repository information, and actual biomarker instance data.

**User manager** implements different roles and appropriate permissions for each user role. It also manages data input/output and updates related to all user information in the system database.

**User interface** is the entry port for end users to interact with BiomRKRS. In the current version, the user interface implements access to all necessary functionalities of BiomRKRS, through appropriate user role and permission management, in a desktop application. The user interface has access to the other system components, namely, user manager, data manager, and query processor. The user interface component also has access to the BiomRKRS core ontology through the ontology management system, mainly to fetch vocabulary lists. At this stage, user queries are formed via direct selection from vocabularies inserted into the graphical components of the user interface instead of through natural language or free keyword-base search mechanisms. Figure 5 illustrates a snapshot of the user interface. A simpler interface has been planned to become available on the World Wide Web with more limited functionality available universally to individual users (subject to consideration of license agreements, especially for external knowledge resources incorporated into BiomRKRS).

## 2.4 System database

To store data pertaining to the different entities in BiomRKRS, a system database has been implemented. These data relate to:

- Ontologies: including all the data related to the physical location of the different external resources imported into the core ontology as well as their update history.
- System: including the physical location of the repository of the system for keeping local copies of related files.
- Biomarker instances: which include the actual biomarker data records. This also includes data on related pathways to each biomarker.
- Users: which includes all data to store for user roles, role permissions, and actual user instances registered in the system.

The system database is only directly accessible to the data manager and user manager components which then make it possible for the other parts of the system to have access to the system database.

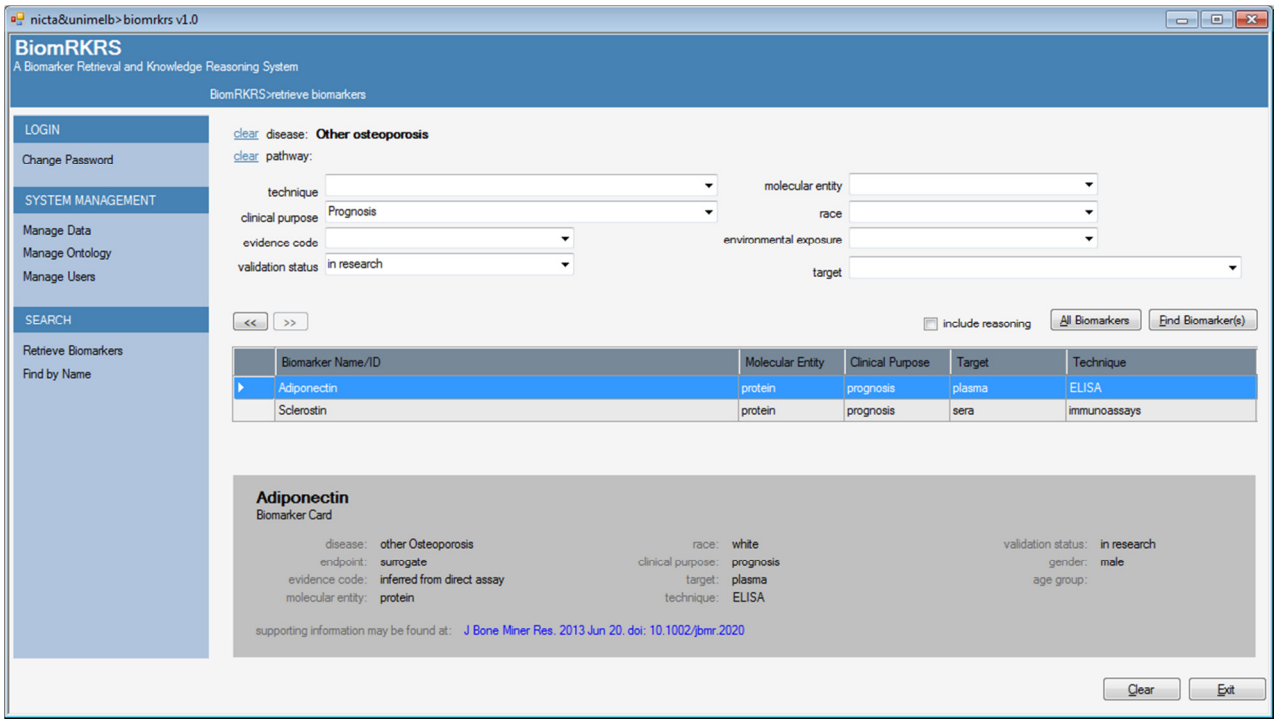


**Figure 5.** A sample screen-shot of the BiomRKRS desktop user interface. User has selected "Other osteoporosis" disease from the list of diseases in the ICD-10 converted ontology, skipped pathway selection tab, and reached to where she can narrow down her biomarker search with some fine-grained biomarker-related contextual attributes including the LOINC measurement technique, clinical purpose, evidence code, validation status, molecular entity, race, and environmental exposure.

## 3 Populating the BiomRKRS database

As mentioned earlier, BiomRKRS, as a biomarker information system, stores general (molecular) biomarker-related information but not patient-specific data. Therefore, there is a need for some mechanism to populate the database with information about identified (and potential) biomarker instances into the BiomRKRS biomarker database. Figure 6 shows the scenarios that have been identified as possible ways for gathering data and feeding them into this database.

The first possibility is to make use of other existing databases that report and store previously identified biomarkers, such as the databases shown on the left side of Figure 6, including OMIM (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005) and PharmGKB (Gong, Owen, Gor, Altman, & Klein, 2008). These individual data sources generally only cover a subset of contextual attributes that BiomRKRS defines as its core concepts and data features (see section 2.1 for the full list of attributes). Hence, BiomRKRS will serve as an integration platform, connecting information from diverse sources together.

The second option for populating the BiomRKRS database is through the use of textual documents in the biomarker-related literature. As we have demonstrated above, PubMed is a significant source of related research articles. In order to use these documents, one may use manual curation to extract and structure biomarker-related information. However, this does not scale to the massive amount of available literature. Therefore, we plan to make use of text mining techniques to automatically extract such information from the text of publications (or abstracts).

At this stage, our BiomRKRS biomarker database includes data that have been manually (by a domain expert) extracted and curated from the published literature related to two specific diseases, i.e., rheumatoid arthritis and osteoporosis.

## 4 BiomRKRS: technical specifications

The first version of the different functionalities of BiomRKRS has been implemented using Microsoft .Net 4.0 standard components. This includes the user interface component as well as all other core and behind-the-scene functionalities discussed in the previous sections.

Database access in BiomRKRS is through Language Integrated Query to Standard Query Language (LINQ to SQL), where the system database itself has been implemented in Microsoft SQL Server. Intermediate wrapper classes have been implemented around all database entities so that any change in the actual means and method of data storage will have a minimal impact on the other data-consumer components of BiomRKRS.

All ontology management and access in BiomRKRS has been implemented using Apache Jena Ontology APIs (Carroll et al., 2004) converted from Java to Microsoft C#.Net. Jena's TDB triple store technology has been used for storing external large ontology files, such as LOINC's instances. The RDFS reasoning schema implemented in Jena has been used to wrap the ontology model constructed using the ontology management component and to create an inference-ready extension of the ontology model. This extended ontology model is then used by the

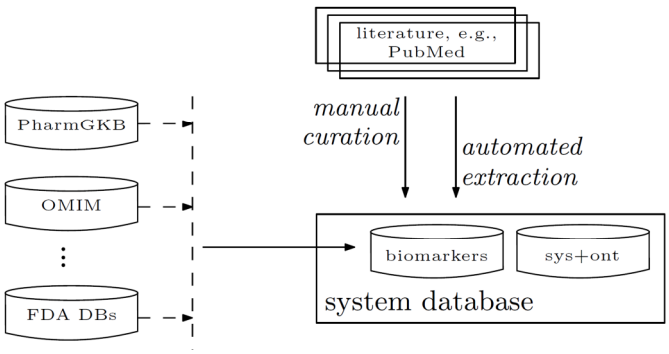


**Figure 6.** Data feeding procedure into BiomRKRS biomarker database

knowledge reasoner component in BiomRKRS for inference purposes.

## 5 A clinical use-case

To demonstrate how BiomRKRS can assist clinical experts with their biomarker information search and related decision making process, we explore an example clinical scenario. In this case, Clinician has some prior and uncertain knowledge about some clinical measurements that have a role as an important prognostic biomarker for disease $d$ that is closely associated with a genetic pathway ($Pw$). Clinician is familiar with the terminology used in the ICD-10 disease classification and has worked with a variety of measurement methods as listed in LOINC's method types. Clinician now decides to search the database of existing and known biomarkers for the following reasons:

- Q1: To understand whether there is any research and/or practical evidence to suggest an important biomarker for prognostic analysis of the genes involved in genetic pathway $Pw$ associated with $d$.
- Q2: If there are biomarkers found in answering Q1, to understand what the current validation status of a particular prognostic biomarker $PB$ is and as a result, whether she needs to carry out any further clinical assays in order to enhance her certainty about the effectiveness of $PB$; also what is the current status of $PB$, in general, for use by other clinicians or researchers.
- Q3: To understand whether further assays are necessary to confirm or enhance the validity of $PB$ being a prognostic biomarker for $d$, what methods of measurement have been used for $PB$ and on what sample (type). This will then lead to making the decision on which measurement method to use from the available methods to Clinician in her lab and how to plan her further study in regards to the subject patients she may have access to.

With these questions in mind, Clinician starts search with BiomRKRS. The process starts by finding the relevant disease term or category from the list of the ICD-10 diseases provided in BiomRKRS. After selecting disease $d$, BiomRKRS shows Clinician the list of associated genetic pathways to $d$. From the list of related pathways, she selects pathway $Pw$ and is now ready to further narrow down her search. BiomRKRS has fetched the list of clinical purposes from the external ontology QIBO (Buckler et al., 2013) and therefore, Clinician can select "prognosis" from the list of available purposes that BiomRKRS offers.

If Clinician knew the exact biomarker she was interested in finding further information about, she would avoid the selection and navigation process by simply searching for the exact title of biomarker $PB$ in BiomRKRS.

Supposing that Clinician does not specifically know any biomarker in this context, to find an answer to her first question (Q1), she now has access to the list of all prognostic biomarkers that relate to disease $d$ and pathway $Pw$ using BiomRKRS. Given that the BiomRKRS list of related biomarkers is not an empty list, Clinician is now able to look at all the evidence related to each data record (each representing a single biomarker) returned by BiomRKRS. Each data record corresponds to certain factors and conditions under which the specific biomarker has been measured, e.g., the patient population (from Experimental Factor Ontology), the measurement method type (from LOINC), molecular entity (from HGNC and HGVS), and validation status of the biomarker. Clinician finds links to the literature where there is evidence that specifically suggest $PB$ be a prognostic marker for $d$ as associated with $Pw$. She may also find some contradictory pieces of evidence that suggest the opposite be true for $PB$. She uses her own judgment, by looking at both sides and available study for each side, to decide whether to accept $PB$ as a biomarker for $d$ or not.

Having decided to accept positive evidence for $PB$ as a prognostic biomarker for $d$, Clinician decides to further her investigation and find an answer for Q2. She can see from the biomarker information record in BiomRKRS that $PB$ has not yet been fully approved by FDA as a qualified biomarker and its current status is "in research". Although there are multiple research publications confirming $PB$ is a prognostic biomarker for $d$, there is no further evidence for it to have been fully investigated and validated against FDA rules and regulations. The validation status "in research" in BiomRKRS confirms that there is a need for more study/analysis in order for $PB$ to become fully "validated" and "qualified". Clinician has now the answer for Q2 and is planning to take her investigation even further.

To answer Q3, Clinician looks at the data retrieved for $PB$, and in particular, the method type feature provided in

BiomRKRS information for $PB$. The fact that BiomRKRS retrieves multiple method types (techniques from LOINC) related to $PB$ makes it possible for Clinician to assess the availability of tools in her lab and program a controlled assay for further investigating $PB$ in the given context. Necessary samples and sample types (mapped to the SNOMED CT specimen concept) for conducting the trials is also a given piece of information from BiomRKRS. Clinician now has access to all information required to carry out further research and validation procedures on $PB$.

The knowledge reasoning feature of BiomRKRS plays an important role in this scenario in the situation where there is no study that has previously shown $PB$ is a prognostic biomarker associated with genetic pathway $Pw$ for disease $d$. BiomRKRS always returns not only the exact matches to Clinician's constructed query, but also a list of biomarkers that semantically relate to her queries. In this case, BiomRKRS retrieves all prognostic biomarkers for $d$ as well as for all diseases $D = \{d' | d' \in [non\_]immediateParent(d)\}$. This is mainly because by reasoning over the ICD-10 ontology, $d$ is-a $d'$. Therefore, any biomarkers found for $d'$ may also apply to $d$. BiomRKRS ranks retrieved data records according to their relevance to Clinician's queries. She will then decide whether to pursue an investigation specifically for $PB$ in the context of disease $d$ if there is evidence that shows $PB$ has been identified as being a prognostic biomarker for a more abstract disease category, i.e., $d'$.

## 6 Planned future work

Development of BiomRKRS is an ongoing task; the current implementation includes all of the components introduced above in basic form. At this stage we have some concrete plans for further extensions and public deployment of the system to clinical experts and researchers. Among these action plans are:

- Implementation of full RDFS reasoning schema, using each inference mechanism for the core concepts from the BiomRKRS core ontology when semantically expanding end user queries.
- Designing and implementing an automated information extraction system for biomarker information extraction from free texts using text mining and natural language processing techniques. Currently all biomarker data in the system database have been extracted manually from related research publications. This text mining component will make it possible for BiomRKRS to store and have access to a large number of biomarker instances (on a larger set of diseases), the information of which can be associated to related textual documents, e.g., PubMed abstracts or full articles.
- Finding genetic mechanisms related to each disease (instead of or) in combination with pathway information. This will give the opportunity to end users to filter out biomarker data based on the participating genes in certain diseases.

## 7 Concluding remarks

With the current state of biological marker (biomarker) information resources and the drastic increase in the number of biomarkers identified for different clinical conditions and purposes, there is a need for effective and efficient information systems that can handle such a large and growing information-base. While there are already a number of systems available for biomarker data management and retrieval, there has been limited reuse of standard and existing vocabularies and none is capable of semantically processing user information requests. Our BiomRKRS system has been designed to overcome these two shortcomings with respect to other biomarker information systems, namely to make use of controlled vocabularies extracted via reuse of other well-established ontologies in the domain as well as to carry out reasoning over the constructed integrated ontology concepts of the system to semantically enhance user queries. As a result, BiomRKRS offers a biomarker data search procedure using controlled vocabulary terms and enables retrieval of exact matches as well as semantically related biomarker data. The semantic reasoning capabilities could potentially support the identification of new indications for previously unknown biomarkers related to certain clinical purposes.

## 8 Acknowledgements

In designing and developing BiomRKRS, we have received great support and knowledge from Professor John Wark at the University of Melbourne and Royal Melbourne Hospital. This work has been supported by National ICT Australia (NICTA) and the University of Melbourne. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

## 9 References

Amplion Research. (2013). BiomarkerBase. Retrieved from http://www.biomarkerbase.com/ Accessed Aug. 10, 2013.

BDW, G. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther*, **69**:89–95.

Buckler, A. J., Ouellette, M., Danagoulian, J., Wernsing, G., Liu, T. T., Savig, E., Suzek, B., Rubin, D. & Paik, D. (2013). Quantitative imaging biomarker ontology (QIBO) for knowledge representation of biomedical imaging biomarkers. *Journal of Digital Imaging*, **26**(4):630–641.

Carrasco, R., & Barton, A. (2010). Biomarkers of outcome in rheumatoid arthritis. *Rheumatology Reports*, **2**(1):26–38.

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., & Wilkinson, K. (2004). Jena: Implementing the semantic web recommendations. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 74–83). New York, NY, USA: ACM. doi:10.1145/1013367.1013381

Cornet, R., & de Deizer, N. (2008). Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making*, **8**:S2.

Crichton, D., Kelly, S., Mattmann, C., Xiao, Q., Hughes, S., Oh, J., Thorniquist, M., Johnsey, D., Srivastava, S., Essermann, L. & Gigbee, W. (2006). A distributed information services architecture to support biomarker discovery in early detection of cancer. In *Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)* (p. 44).

Evolvus. (2013). Evolvus biomarkers database. Retrieved from http://www.evolvus.com/Knowledge Management/Databases /BiomarkersDatabase.html/ Accessed Aug. 10, 2013.

Forrey, A., McDonald, C., Huff, G. D. S., Leavelle, D., Leland, D., Fiers, T., Charles, L., Griffin, B., Stalling, F., Tullis, A., Hutchins, K. & J, B. (1996). Logical observation identifier names and codes (LOINC) database: A public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry*, **42**(1):81–90.

Gong, L., Owen, R. P., Gor, W., Altman, R. B., & Klein, T. E. (2008). PharmGKB: an integrated resource of pharmacogenomic data and knowledge. *Curr Protoc Bioinformatics*, **14**(14.7).

Gray, K. A., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W., & Bruford, E. A. (2013). Genenames.org: the HGNC resources in 2013. *Nucleic Acids Research*, **41**:D545–52.

GVK Biosciences. (2013). GVK BIO online biomarker database. Retrieved from https://gobiomdb.com/gobiom/ Accessed Aug. 10, 2013.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, **33**(suppl 1):D514–D517.

Human Genome Variation Society. (2013). HGVS: Human Genome Variation Society. Retrieved from http://www.hgvs.org/ Accessed Aug. 10, 2013.

International Council Of Nurses. (2013). International classification for nursing practice. Retrieved from http://www.icn.ch/pillarsprograms/international-classification-for-nursing-practice-icnpr/ Accessed Aug. 10, 2013.

Liatris Biosciences LLP. (2013). Biomarker databases. Retrieved from http://www.liatrisbio.com/biomarker_databases.html Accessed Aug. 10, 2013.

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A. & Parkinson, H. (2010). Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**(8):1112–1118.

Moller, M., Sonntag, D., & Ernst, P. (2010). Modeling the international classification of diseases (ICD-10) in OWL. In A. Fred, J. L. G. Dietz, K. Liu, & J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management: Second International Joint Conference, IC3K* (pp. 226–240).

Olson, S., Robinson, S., & Giffin, R. (2009). *Accelerating the development of biomarkers for drug safety: Workshop summary*. The National Academies Press.

SciClips. (2013). Cancer biomarker database. Retrieved from http://www.sciclips.com/sciclips/diagnostic-prognostic-cancer-biomarker-main.do/ Accessed Aug. 10, 2013.

Szulc, P., & Delmas, P. D. (2008). Biochemical markers of bone turnover: Potential use in the investigation and management of postmenopausal osteoporosis. *Osteoporosis International*, **19**:1683–1704.

The Gene Ontology Consortium. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genet*, **25**:25–29.

Thomson Reuters. (2013). Biomarkers module of Thomson Reuters Integrity. Retrieved from http://thomsonreuters.com/products/ip-science/04041/biomarkersmodule-cfs-en.pdf/ Accessed Aug. 10, 2013.

Vasan, R. S. (2006). Biomarkers of cardiovascular disease: Molecular basis and practical considerations. *Circulation*, **113**:2335–2362.

Vasikaran, S., Eastell, R., Bruyre, O., Foldes, A., Garnero, P., Griesmacher, A., McClung, M., Morris, H., Silverman, S., Trenti, T., Wahl, D. A., Cooper, C. & Kanis, J. (2011). Markers of bone turnover for the prediction of fracture risk and monitoring of osteoporosis treatment: A need for international reference standards. *Osteoporosis International*, **22**(2):391–420.

Weigel, M. T., & Dowsett, M. (2010). Current and emerging biomarkers in breast cancer: Prognosis and prediction. *Endocrine-Related Cancer*, **17**(4):R245–62.

World Health Organization. (1992). *ICD 10: International Statistical Classification of Diseases and Related Health Problems Volume 1*. World Health Organization.

Yang, I., Ryu, C., Cho, K., Kim, J., Ong, S., Mitchell, W., Kim, B., Oh, H. & Kim, K. (2008). IDBD: Infectious disease biomarker database. *Nucleic Acids Research*, **36**:D455–60.

Younesi, E., Toldo, L., Mller, B., Friedrich, C. M., Novac, N., Scheer, A., Hofmann-Apitius, M. & Fluck, J. (2012). Mining biomarker information in biomedical literature. *BMC Medical Informatics and Decision Making*, **12**:148.