12

Multi-Faceted Search and Navigation of Biological Databases

Mahoui M., Oklak M. and Perumal N.
Indiana University School of Informatics, Indianapolis, USA

1. Introduction

1.1 Challenges in bioinformatics data integration

The field of biology has clearly emerged as a data intensive domain. As such, several challenges facing the design and integration systems for biological data exist [1] and continue to persist [2] despite the efforts of the bioinformatics community to reduce their impact. These challenges include 1) the large number of available databases, 2) their often http/HTML based mode of access, 3) their syntactic and semantic heterogeneity. The challenges are strongly supported by the number of increasing databases publically available—varying from 96 databases in 2001 to more than 1,330 in 2011 [3]. The available databases cover different data types including nucleotide databases such as GenBank [4], protein databases such as Uniprot [5], and 3D protein structure databases such as PDB [6]. While the majority of available secondary and tertiary databases are derived from primary databases such as PDB or Swissprot [7], and therefore contain redundant data, they generally provide the research community with added features resultant from studies conducted by the database providers.

Parallel to the exponential increase in volume and diversity of available data, there has been an exponential increase in querying these databases as a routine task when conducting research in biology. Retrieved data is often integrated with other data produced from remote or local sources and/or manipulated using analytical tools. Consider, for example, the study of genes associated with a particular biological process or structure. An isolated DNA sequence would be screened against known gene sequences in GenBank, converted to a putative protein sequence and screened against SwissProt. Finally, any region showing similarity to a known gene or protein can then be queried for known 3D structures and be visualized using the PDB database to obtain a general idea of putative structure and function of a newly isolated gene. A subsequent search of various specialized databases would still be necessary to obtain up-to-date information regarding analogous research in other model organisms and associated pathway structures. To support the types of studies involving multiple biological databases, several integration systems have been proposed [8-13]. To characterize the existing systems several dimensions have been proposed [1, 2], including the aim of integration and the integration approach. When analyzing the aim of integration, the existing systems can be largely classified as either portals-oriented or query-oriented. Portals-oriented systems have their focus on providing an integrated view to the accessed databases, where notable examples include SRS [14] and NCBI Entrez [15]. Query-

oriented systems, focus on supporting user queries that can span more than one database. Examples include TAMBIS [9], BACIIS [12] and Biomediator [16]; and to some extent workflow systems such as Taverna [17]. With respect to data integration approaches, three main alternatives have been deployed: data warehouse, data linkage, and wrapper-mediator.

In the wrapper-mediator approach, the integrated data is not physically stored at the integration system as it is in the warehouse approach. Rather, it is obtained at the time of the query using the wrappers to interface with the data sources and the mediator to generate a uniform view of the data for the integration system. This principal advantage of the mediator approach is that it fits very well with the ever growing number of databases and their short life expectancy [2].

1.2 Moving the data search into the data systems view

The data search behavior of pre-genomics era researchers was largely a one-gene-at-a-time approach. Indeed, transitioning from wet-lab experiments progressively towards more in-silico experiments, post-genomics researchers will often start from an incomplete biological entity, such as the DNA sequence, and use available databases to annotate the entity with multiple biological features (or facets) to build a more comprehensive perspective. To address these types of queries, current databases and the majority of existing portal systems typically provide users with a keyword search, where results are given as a list of top-ranked records that match the query. Clicking on, or selecting, any record will retrieve additional annotated information about the target record including references to other databases. This record-based approach is clearly not scalable when considering the number of returned records from databases, especially with portals integrating several complementary databases. Systems such as GeneCards [18] are closer to providing users with a more comprehensive view of the records without having to search for other databases (in addition to other options such as advanced search and output parameters). However, the record-based approach requires the user to "click" on each record sequentially to progress through the rest of the features (facets) and to manually compare returned records.

High-throughput technologies and advances in next-generation sequencing have placed an increasing emphasis on the need for a systems level approach to the study of the life sciences, with the generation of hundreds of thousands of genomic and proteomic data points rather than only a few hundreds. Concurrent with these developments, there is an increasing need to perform bioinformatics studies at this systems level, as well as the gene level. For example, a protein such as Notch1 which is involved in lymphocyte development acting at the cell surface, could be the starting point for searches on associated signaling and metabolic pathways, protein-protein interactions, transcriptional regulatory networks, and drug targets important in this system. A holistic systems level search will provide the geneticist or developmental biologist a clear an advantage in terms of time, effort, and knowledge gain, previously unattainable by record-based searches. Specific applications exploring the relationships between biological entities such as protein-protein interactions, e.g., the DIP database [19], already provide a systems view. Biological databases and database portals are currently lacking in this pivotal capability. A faceted classification approach provides a multi-dimensional view of the data that can be used to both group and aggregate the data. Similar to the OLAP approach and data cube technology [20], biological data can be represented by a set of biological features or facets (i.e. dimensions) such as gene

information, pathway information, drug targets information, etc. These facets can in turn be used to conduct an interactive, discovery-driven search where the user can navigate through the multi-dimensional data, refining the search by drilling down or rolling-up any hierarchical facet and/or by combining multiple facets.

1.3 A multi-faceted data integration approach for querying biological databases

We propose Biofacets, a multi-faceted data integration system for querying biological databases. The key feature of Biofacets is the support of multi-faceted searching/browsing of biological databases, thus providing a true representation of the system view of biological data. Biofacets is based on a wrapper approach where search queries submitted to Biofacets are relayed to the integrated biological databases, and results are aggregated on the fly using the multi-faceted scheme.

The main contribution of the paper encompasses the following:

- Demonstrate the potential of multi-faceted paradigm in advancing biomedical research.
- Understand the challenges that surround the building of wrapper-based multi-faceted data integration system for biological databases.
- Describe the solution we propose to address these challenges. Specifically, we describe the evolution of Biofacets architecture that led to a more scalable and reliable infrastructure.

2. Related work

2.1 Data integration of biological databases

While the focus of Biofacets is to primarily empower biological databases with faceted searching/browsing, data integration issues are closely linked to the project. As described in section 1, several integration systems have been proposed in the bioinformatics community (see [2] for a recent survey). Integration Systems vary from simple but powerful settled-warehouse solutions to more flexible ones using technologies such as mashups that expose the researchers to a greater control and therefore more apriori informatics knowledge in resolving the integration issues. Recently, hybrid solutions [21] involving the semantic web and the wrapper-mediator integration approach (also known as view integration) have provided a step forward towards leveraging the flexibility of the available integration architectures while reducing the impact of the semantic heterogeneity that characterizes biological databases. Note though, we have yet to see the impact of new paradigms such as dataspace systems [22, 23] that offer a less rigid but perhaps more expandable integration architecture in designing new biological data integration systems.

Biofacets uses a wrapper mediated approach on Local As View (LAV) data model approach as opposed to a Gloabl As View (GAV) approach [24]. This approach is particularly flexible for data sources that are less stable as is the case for biological databases (see section 3 for more details). Another feature of the Biofacets data integration approach is that, as a portal, the mapping between the global schema and the source schema is straigtforward and the emphasis is on mapping the source schema into the global schema.

2.2 Faceted browsing

Faceted searching, the main motivation behind building Biofacets, is less explored in bioinformatics despite its popularity in other applications and in the research community

[20, 25, 26]. The majority of research effort providing automatic support for faceted data search is related to (a) the automatic generation of the facets and their hierarchies (hereafter referred to as the ***faceted scheme***), and (b) the design of the faceted user interface. Very little published work is dedicated to the implementation details describing (c) how the facet scheme is to be deployed within a collection; that is, how facets are assigned to records/documents and how the facet values are extracted during query time.

a. *Automation of the faceted scheme:* Before faceted search became a popular topic, several research contributions have been described in the area of document clustering [27-31]. For example, the Scatter/Gather [27] algorithm is based on a recursive version of the agglomerative clustering algorithm. The advantage of clustering is that it is an unsupervised technique. The main criticism addressed to this class of work is that the clustering-based approaches generate a set of features (keywords) as opposed to producing a representative label for each cluster. This method makes their deployment for faceted search not straightforward. Another approach [32-34] aims at generating hierarchies of terms to support data search/browsing. The subsumption method is proposed in [32], whereby a term "x" is said to subsume term "y" if $P(x/y) \geq 0.8$ and $P(y/x) < 1$. The subsumption relationship is also utilized in [33] where the main contribution is the expansion of the collection terms with external resources such as Wikipedia and Yahoo terms in addition to identifying named entities to help identify the main facets. The automatic method proposed in [34] makes use of hypernyms on WordNet's synsets, together with a hierarchy minimization method to generate the hierarchical scheme.
b. *Design of the faceted interface*: This aspect has drawn the attention of many research works [35-40], especially the work led by Heart et al. Usability studies [37] were conducted and several guidelines on the design of the faceted interface were described and implemented in the Flamenco Project [41]. These guidelines include availability of aggregate counts at each facet level and combination of facets during refinement. Flamenco intentionally exposes the metadata associated with the images in its database to allow users to navigate along conceptual dimensions or facets describing the images. Software such as FacetMap [32] provides automated tools to develop faceted classification systems. However, it assumes the availability of both data and metadata (i.e. facets scheme) to build the faceted interface. Note that other work [39, 40] displayed the data as two dimensional tables to correlate between facets.
c. *Mapping between facets and documents/records:* Previous work [42] provides a good description of the internal documents and data representation needed to support the faceted classification. They assume that the mapping of the facets to documents is available and that each facet is available as a path of labels in the hierarchical scheme. A modified inverted index together with a forest of facets hierarchies is used to match the query (i.e. keyword with searched facet) to the documents and build their faceted view including the counts at each facet level. They also provide the users with the ability to perform aggregate functions in addition to count, a feature that is very suitable for business intelligence.

In Biofacets, the browsing scheme serves as the global schema for the wrapper-mediator data model. Moreover, in the current version of Biofacets, the scheme is generated manually as the main current focus is to showcase how multi-faceted browsing can be leveraged when searching biological databases.

3. Biofacets design

3.1 Biofacets architecture

Biofacets is designed as a client server application to be used as an enhanced portal between researchers and the wealth of databases publicly available in the Web. Figure 1 highlights the various modules of the Biofacets system and their current status in the design/implementation process [43-45]. The user query is forwarded to the Query Module, which in turn passes it to the Cache Management Module, to determine whether the query has already been cached; in which case the results' URLs are immediately available. In case the query is not cached, it is processed by the Query Module. A keyword search is launched against each integrated database using the source information from the Source Knowledgebase. As results become available from each database, they are passed on to the Faceted Classification Module, which assigns facet values to each record using the Facet Knowledgebase. Finally, the data records, together with the corresponding facet values, are passed on to the Presentation Module, which prepares a presentation file to be viewed via the Web Interface. Note that the results are grouped by facets and no specific ranking is used to list them within a facet.

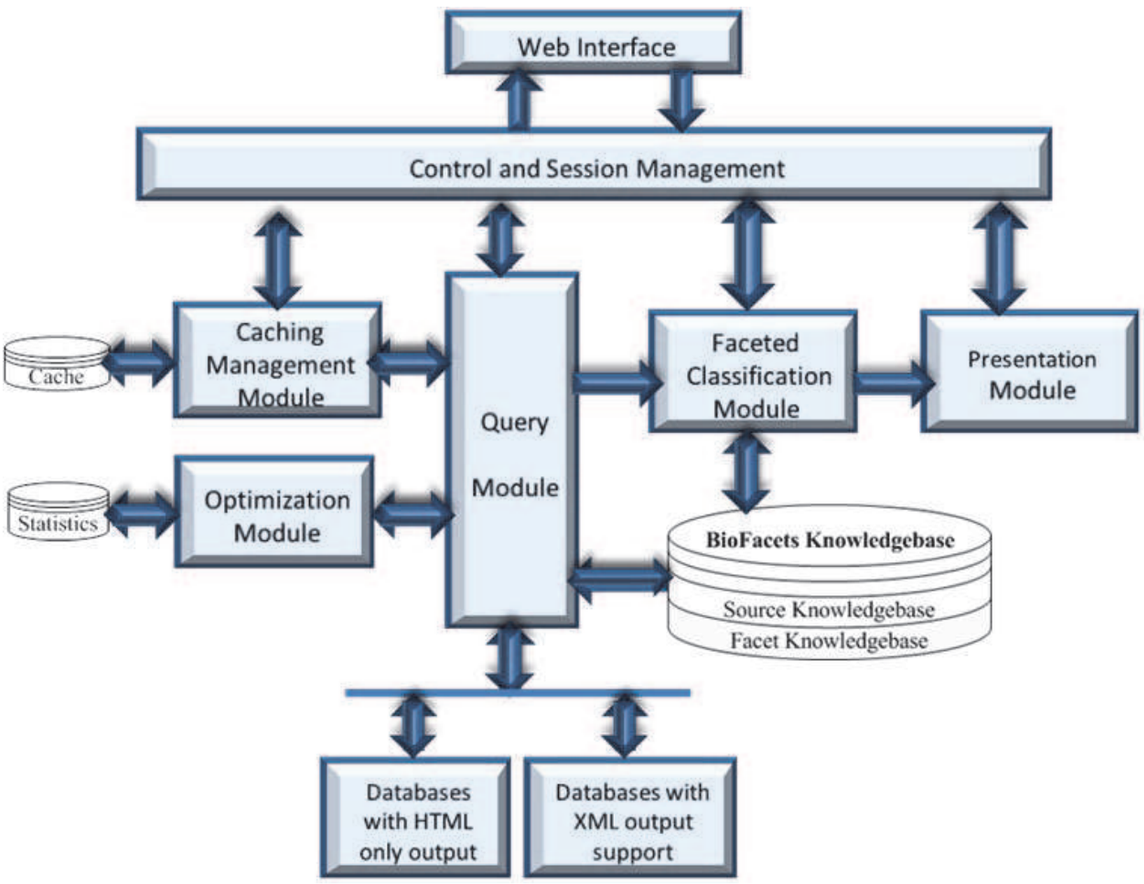

Fig. 1. Overall Architecture of Biofacets

In the following sections we will detail the core modules essential to Biofacets.

3.2 Wrapper-based integration system for searching remote biological databases

Biofacets is both a meta-search engine and an integration system. Results retrieved from the databases can be integrated into a uniform internal representation, thus resolving the heterogeneity issue characterizing biological databases. More precisely, the role of the wrapper is to ensure (i) querying of the supported databases, (ii) extraction of data from retrieved results pages, and (iii) integration of results using a shared terminology into an internal representation. The last two tasks are performed together, though they are two distinct processes.

To perform the data integration phase we distinguish between two types of databases: databases that rely only on http-html protocols to make available their data, and databases that support XML as an option for results output. Within the latter group we find databases that provide XML as an output in addition to the HTML support, and databases that provide support for web APIs to query their data with XML as one of the options for output. Next we will describe the wrapper solution for each of these two types of databases.

3.2.1 Databases with no support of XML output

Most of the web-based biological databases are only accessed through http protocol using a web interface requiring integration systems to mimic user search behavior to query them. Biofacets stores the base URL for wrapper use as part of the database schemas in the *source knowledge base*. The wrapper uses the base URL with user search terms to send the search query. The query results are generally available as html pages with a mix of data and html tags. Extraction rules are necessary to the process of extracting from the HTML pages the data that identify the biological entity (e.g. organism name) and its value (e.g. "Drosophila Hydei"). The first version of Biofacets uses an extended version of HLRT rules [46] for data extraction. The main principle of HLRT rules is the identification of landmarks from which to precisely extract the value of the identified labels. The landmarks located left of the target value are known as "Head" and "Left" delimiters, and those located to the right are known are "Tail" and "Right" delimiters. The wrapper engine uses extraction rules for extracting entities and their values from both summary and extended pages; where summary pages usually include summary information for each record retrieved, and extended pages provide detailed information for one record. The wrapper will use the schema defined for each database to generate the internal representation (both summary and extended) of the results, serialized in XML, to be used by the faceted classification and the presentation modules (Figures 2 and 3).

```
<field name="record">
  <extraction_rules>
    <ld><b>+1:+</b></ld>
    <rd></rd>
  </extraction_rules>
  <field name="protein_definition"
         save_value="true">
    <extraction_rules>
      <ld>DEFINITION</ld>
      <rd>ACCESSION</rd>
    </extraction_rules>
  </field>
</field>
```

Fig. 2. Sample Summary extraction rules

Note that the entity labels (e.g. protein_definition, ncbi_protein_identifier) used to generate the internal representation of the results are part of the facet knowledgebase used to integrate the results of queries resulting from different and heterogeneous databases.
Within the first version of Biofacets the database schema (Figure 2) was manually generated. Currently we are working on providing automation support to the process of data extraction and data labeling (see Section 4).
Databases schemas include the information necessary to query the database (i.e. the base URL) and to extract the facets and facet values of the labels providing the uniform view of the integrated data, in addition to HLRT rules. These labels are part of the Biofacets knowledgebase (see section 4).

```
<record
complete_url="http:/www.ncbi.nlm.nih.gov/entrez/
    viewer.fcgi?db=protein&amp;val=7436"
    datasource="Entrez Protein">
   <extended_record_link>
     <value>

/entrez/viewer.fcgi?db=protein&amp;val=7436
      </value>
    </extended_record_link>
    <ncbi_protein_identifier>
      <value>CAA36808</value>
    </ncbi_protein_identifier>
    <protein_name>
      <value>histone H2b</value>
    </protein_name>
    <organism>
      <value>Drosophila hydei</value>
    </organism>
</record>
```

Fig. 3. Sample Summary extraction rules

3.2.2 Databases with support of XML output

The majority of biological databases offering support for XML output are from NCBI [47] and EBI [48]. Both provide access to a large number of databases using (i) APIs to facilitate the querying of databases and/or (ii) an XML representation of query results. For example NCBI Entrez makes available Esearch and Efetch utilities [47].
When dealing with this type of databases, querying still requires URL submission. However, writing extraction rules is reduced to writing XSLT transformation rules [49]; a standard process as compared to custom HLRT rules.
While the XML presentation option is increasing in availability for results presentation, mapping is still required between database-specific entity names (i.e. XML element names/attributes) provided by the database XML output result and the internal labels used by the internal XML result presentation as provided by the facet knowledgebase (for integration purposes).

3.3 A facet-based data model for results integration

The main feature of the Biofacets system is the proposal of a dynamic, hierarchical, and faceted classification approach that supports the categorization of query results by dynamically assigning facets to retrieved data records. The main difference between a static facet approach and a dynamic approach lies in the fact that for a static approach, the assignment of facets to data items is statically performed *a priori* before the faceted system is deployed. This assignment uses either metadata information provided with the data or the expertise of professionals. This is the case of the faceted systems supporting commercial Web sites such as Amazon.com. On the other hand a dynamic faceted scheme is deployed on the fly to assign facets to retrieved results. Therefore, the specification of a dynamic faceted classification approach includes determining the methods by which facets are assigned to each data item.

3.3.1 Specification of the faceted classification model

A facet is simply a method of classification. It groups together results with the same value for a particular category or field and provides a view of the result set classified according to each of these categories. The categories defined are mutually exclusive and hence facets are orthogonal. Using faceted classification, a record is described by combining facet values. In Amazon.com a subset of the facets used to describe clothes, for example, are price, brand and size.

We define a facet using three criteria: (i) its *depth*, (ii) its *depth generation,* (iii) and its *value assignment*. With the first criterion a facet can be either *flat* such as the "Color" facet, or *hierarchical* such as the "Location" facet. The "Location" facet is hierarchical as it can be broken down into the "Country" facet, then into "State/Province" facet; and finally into the "City" facet.

Regarding the assignment of values to facets, we identify two approaches: *static* or *dynamic*. Static facets are facets for which the value assigned to a record is determined without the knowledge of the record; usually using the information about the database to which the record belongs. For example, the static facet "Data Type" will take a fixed value from the predefined set (e.g. {protein, gene, literature}). On the other hand, a dynamic facet is a facet for which the value assigned to a record depends on the record value. In this context and based on a comprehensive survey of a large set of biological databases, we identified two main methods by which values are assigned to facets. More precisely, the value of a facet is either directly available within the targeted record or indirectly obtained using the information provided by the record. In the latter case, the facet value is extracted from a third party database. This has led to the specification of three types of classification rules for facet value assignment viz. the *fixed value*, *field value* and *lookup value* rules. The fixed value rule is used with static facets and assigns a predefined value to each record belonging to a particular database. The *field value* and *lookup value* rules are used with dynamic facets. The *field value* rule assigns the value of a field in a record as the facet value for that particular record, while the *lookup value* rule does query another database to obtain the facet value.

Facet depth generation concerns hierarchical facets and specifies whether the hierarchy of the facet is known a priori before it is deployed by the classification process or it is dynamically generated during the classification process. The need for dynamically generating a facet hierarchy is proposed to take into account large exhaustive hierarchies such as the organism hierarchy for which only a subset is generally needed for a query. Moreover this hierarchy is developed and maintained by third party organizations, such as Newt [50] and NCBI for organism facet. For dynamically generated facet hierarchies, the classification rule is a combination of *lookup value rule* and the *depth parameter*. The lookup

value rule is used to obtain the partial tree locating a record in the facet hierarchy, and the depth parameter is used during the dynamic hierarchy generation to specify the depth of the generated hierarchy obtained by combining the partial trees of the records.
Assignment of facet values to records is decided at the data source level. Thus, records from the same data source will share the same set of facets; each facet is assigned using the same rule. Therefore, for each database supported by Biofacets, one needs to specify the set of rules that apply to the data source, and the instantiation of the facet rule. For a static facet (e.g. "Data Type"), the static value is specified (e.g. "protein type" for NCBI Entrez (Protein)). For a dynamic facet, we specify the type of rule applied, as well as the fields or the third party data sources involved (figure 4).

```
<facet fname="literature" facet_value_range="dynamic" isHiearchicalDynamic="false">
    <facet fname="authors" facet_value_range="dynamic" isHiearchicalDynamic="false">
        <terminal_non_hierarchical_dynamic_node>
            <database dname="pubmed" classification_method="field_value">
                <classification_rule>
                    <field_value_computation_rule>
                        <field>authors</field>
                    </field_value_computation_rule>
                                                </classification_rule>
                                    </database>
```

Fig. 4. Faceted classification specification extract

The set of classification rules that assigns facet values to each facet, for each database, is referenced hereafter as the *database_facets_mapping*. Facets hierarchy (or faceted scheme) and *database_facets_mapping* are serialized using XML. XML schema is used to specify the structure of a faceted scheme and classification rules supported by Biofacets. A facet specification (Figure 4) includes its name (e.g. literature), its facet type (i.e. static, dynamic), whether it is hierarchical or not, and whether its hierarchy is dynamic or not. Each facet type is specified by the facet classification rule(s) that can be used to extract facet values from data sources. XML schemas are used to validate a new facet or a new database added to the Biofacets system. Facets' specification and the *database_facets_mapping* (XML instance and XML schema) compose the *facets schemas*, which is part of the *Facets Knowledgebase* (see Figure 1). Note that while facets schema specifies the structure of the facets and databases classification rules, the design of the faceted schema itself is a separate task part of the designing of the Biofacets' ontology.

3.3.2 Assigning facets to data records

The algorithm we propose for assignment of facet values described in [51] uses *database_facets_mapping* to assign facet values to the set of records for the specified facet, for each database. The faceted classification module receives a set of records extracted from the summary result page. If the type of facet is static, the corresponding value is extracted from the *database_facets_mapping* file and assigned to all records. If the facet is dynamic, each record in the summary information is processed. More precisely, if the classification method is *field value*, the field specified in the *database_facets_mapping* is searched in the extracted summary information. If present, its value is assigned to the record for the specified facet. Otherwise, the extended extraction rules are applied in an attempt to find the field. If the field is not found in both, a value of "undefined" is assigned. In case of the *lookup value* method, the record with a

given value for the lookup field is searched for in the third party database (both the lookup field and the third party database are specified in *database_facets_mapping*). Once the record is located, a facet value is assigned similarly to the field value method.
Note that for the databases with support of XML output, the summary XML pages contain only the identifiers of records that satisfy the search query. The information to be used by Biofacets is in the extended XML pages.

3.4 Faceted classification for data querying and result browsing

Faceted classification can be used to support researchers (1) in browsing the results returned by the integrated databases; (2) and in targeting the search (i.e. advanced) query; with the ability to specify a set of values for a given facet at the time the query is submitted; for example, searching within the facet protein name for records with protein name "tyr". These values submitted to guide the search will then be used to filter out the results before they are displayed to the user. While the first goal is overall supported by the current prototype (Figures 5-7), the second goal is supported to the extent that researchers can specify the facet they are interested to find data records about.

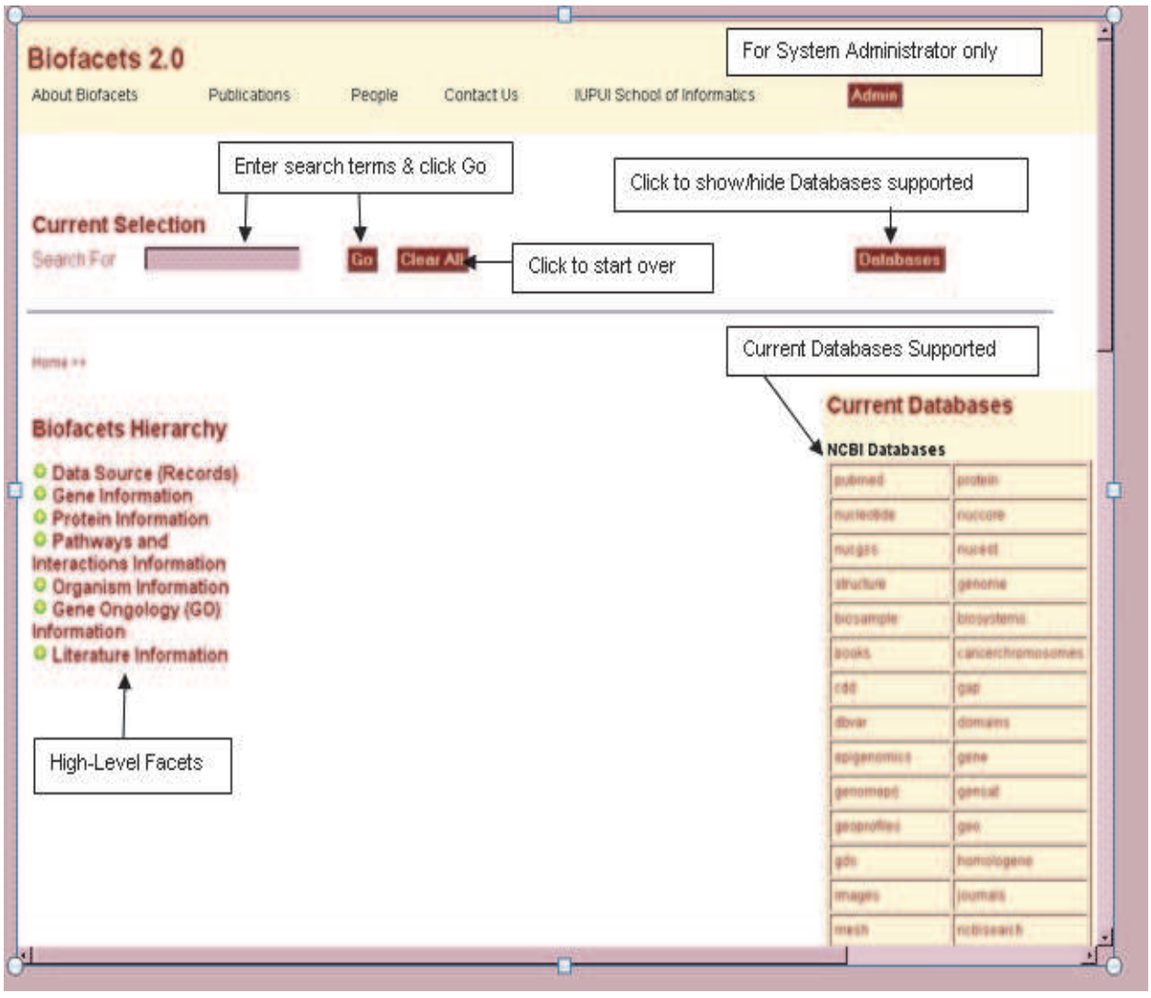

Fig. 5. Biofacets Main Entry Page

With keyword search, users can specify which facets they want results to be grouped by. Once the results are displayed, the user can refine them by zooming-in (specialization) or zooming-out (generalization) in the facet hierarchy. As part of the refinement, the user can also select another main facet to narrow down the results using a combination of facets. To facilitate the process of searching and refinement, we incorporate state-of-the-art guidelines into building Biofacets' interface [52]. This include features such as the display of the record count at each level of the facet hierarchy, and the indication of the list of the facets involved in the current displayed results with the bread crumb technique.

The main entry page to Biofacets (Figure 5) includes information about the main facets supported by the integration system, and the databases currently supported[1], in addition to standard information such as contact information list of publications, etc.

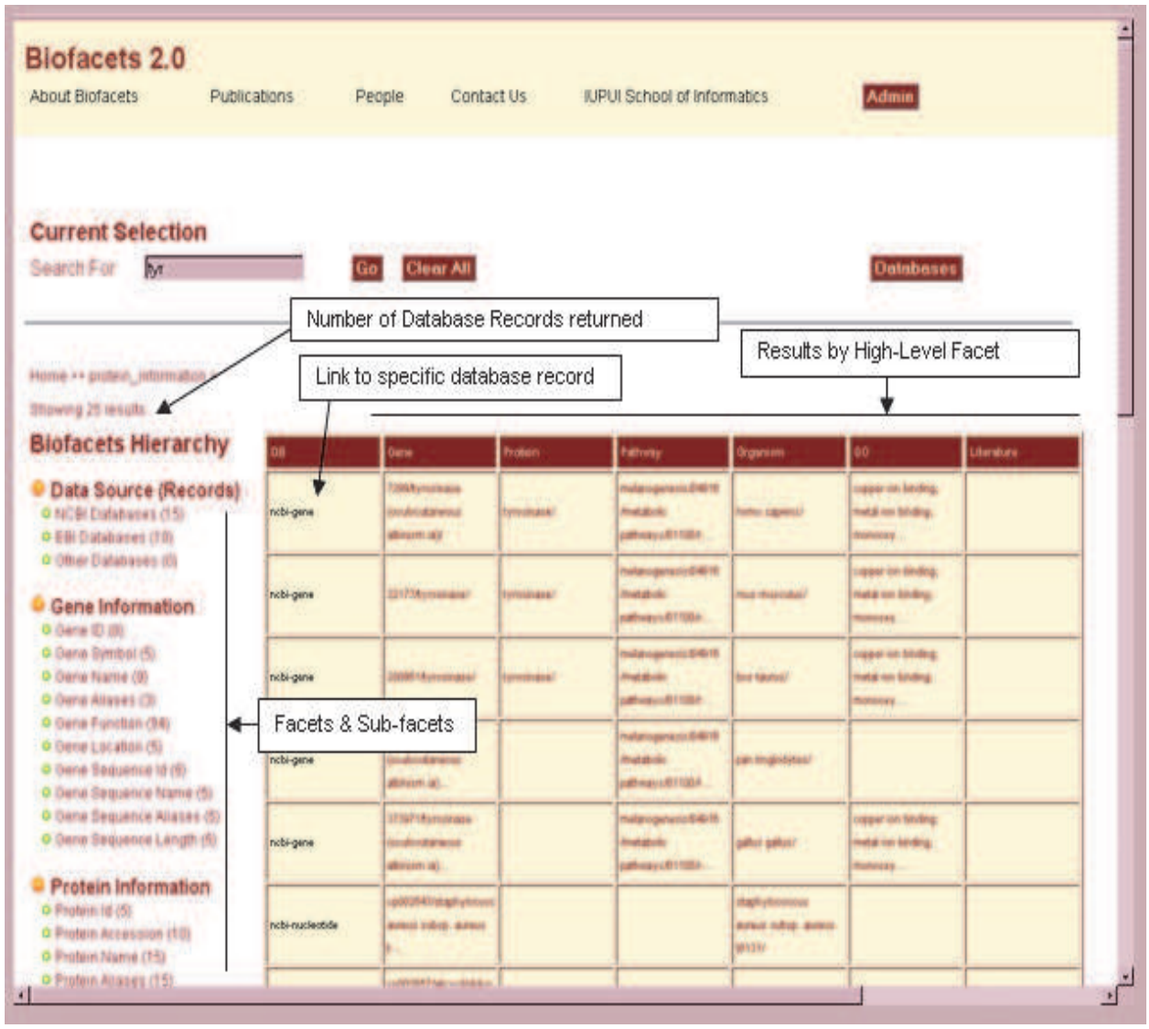

Fig. 6. Results Returned

[1] At the time the screenshots were taken only databases that support XML output are searched as the Biofacets system is currently in the process of redesigning its component that handles databases with no XML support.

Figure 6, depicts the results returned by the search for records related to "tyr". The facets data source, protein information and gene information are expanded to highlight some of their sub-facets. To each (sub) facet the number of records for which the facet has a value is displayed. The records matching are summarized using a table. This summarization technique is becoming very popular with biological databases. We choose the following facets to summarize the content of the records: database name, gene name, protein name, pathway ID, organism name, gene ontology term, and literature pubmed ID. Links to the original records are also provided for each record.

Fig. 7. NCBI Database Results

In figure 7, the user clicks on NCBI databases facet and the initial results are filtered using this facet. Only the records corresponding to NCBI databases are displayed in the main frame of the results page. Figure 7 also shows the progression of the bread crumb option to help the user keep track of the filtering process he/she is performing. Note that the bread crumb option can also be used to zoom-in and zoom-out in the results.

Part of the future work is to conduct an evaluation and validation of Biofacets' browsing interface in order to ensure that Biofacets is tailored to researchers searching and browsing needs.

3.5 Biofacets knowledgebase

Biofacets knowledgebase is the backbone for Biofacets system. It includes (1) the source knowledgebase deployed by the query module, and is composed of the schema of the integrated databases; (2) the facet knowledgebase composed of the faceted scheme and its formal description, *facets schemas*, which is used by the classification module; (3) and Biofacets ontology used as a common terminology by both the wrapper to reconcile between the heterogeneity of the integrated data, and by the classification to support the vocabulary used by the faceted scheme. While the faceted scheme vocabulary is part of the Biofacets ontology, *its structure is not a subset* of the Biofacets ontology; the main reason being that the faceted scheme is used for results browsing and query refinement while the Biofacets ontology is used as an internal representation data model.

3.5.1 Biofacets ontology design

An ontology is the specification of a conceptualization as it consists of a set of concepts expressed by using a controlled vocabulary and the relationships among these concepts, which are used to infer the meanings of these concepts. In bioinformatics, ontologies are becoming popular data models. They can be classified, according to their use, into three categories: domain-specific, task-oriented, and general [53]. An example of a domain-specific ontology includes Gene Onotology GO [54]. Examples of task-oriented ontologies include EcoCyc [55], TAMBIS [9] and BACIIS [56]. Biofacets ontology falls into this category as its purpose is to facilitate the task of categorizing data records. More precisely, Biofacets ontology was designed to satisfy the following:

- Provide a shared terminology to allow mapping between databases' specific terms by having them correspond to unique terms provided by the terminology
- Provide support for the hierarchical structure that characterizes faceted classification schemes
- Provide support for other relationships between concepts in addition of the parent-child relationship.

While the first two conditions can be provided by a general taxonomy, the third condition requires the use of ontologies to represent more than subsumption relations between concepts. Provision for such relationships is important to support automatic assignment of facets to databases (see section 4).

In addition to including concepts in biology domains (e.g. DNA sequence), concepts related to bioinformatics (e.g. id of a protein) also need to be represented in the ontology. Moreover, general concepts such as those related to disease or literature information are also part of the shared vocabulary. Task-oriented ontologies such as Mygrid [57] and SIBIOS [58] are too complex for the purposes of Biofacets, as these ontologies are designed to support in-silico experiments, deploying both databases and analytical tools such as NCBI Blastn [59]. Leveraging on our experiences building BACIIS [56] and SIBIOS [58] ontologies, we adopted an incremental design of Biofacets ontology. More precisely, the purpose was not to provide a comprehensive ontology that will support all potential databases before starting to use Biofacets, but rather to provide an ***ontology structure*** that can be easily updated with new

concepts. Toward this objective, we combined the following research approaches to build the core Biofacets ontology:

- Surveying ontologies: this includes not only standard ontologies such as GO ontology and Mesh ontology, but also task specific ontologies such as TAMBIS and Mygrid ontologies
- Utilizing popular categorizations such as the categorization supporting the nucleic acid research collection [3] and DBCat categorization [60]. This will provide insight with respect to the hierarchical structure of the ontology and the concepts names to be used
- Initializing the integration process with popular databases such as UniProt [61] and data centers such as NCBI. The aim is to leverage on the popularity of these databases and utilize as much as possible of their terminologies when defining Biofacets ontology terms.

3.5.2 Biofacets faceted scheme design

The current Biofacets portal is supported by a manually generated faceted scheme. The design is based on the study of a list of the 25 most popular databases specializing in different topics selected from the Nucleic Acid Research (NAR) database collection [62]. The main facets identified in the study are "data-type, data-source, literature, protein-info, gene-info, organism-hierarchical". Each main facet contains up to 3 hierarchy levels including the facet values. The facet "data-type" groups the results based on the type of the data described in the record (e.g. protein, gene, literature, alternative splicing). "Data-source" facet has two sub-facets: "NCBI-databases" and "other-databases". EBI databases facet was added at a later stage. The facet "hierarchical organism", grouping records according to their lineage information, is special in the sense that facet hierarchy is not stored locally; but it is generated dynamically by integrating the facet paths provided by each record[2]. The facets Pathway and interaction information as well as Gene ontology information where added as later stage. The total of facets currently available for researchers is 33 facets.

3.6 Biofacets performance and cache management

Biofacets is designed as a meta-search engine for biological databases enhanced with a classification mechanism of queried results. Two main factors pose a bottleneck for the overall query response time: (1) the time necessary to query remote databases and get the results back and (2) the time necessary to classify the results due to the dynamic nature of the faceted classification approach.

In the domain of biology, indexing biological data seems inappropriate purely due to its sheer volume and heterogeneity; which makes the prediction of user queries an unpractical task. To reduce the impact of these factors, the solution we propose consists of (1) caching the query results, especially the most frequent queries and (2) querying all supported databases in parallel, while progressively providing the results to the user as soon as they become available.

The main role of a cache management component is to ensure efficient retrieving/storing of results from/into the cache, and appropriate cache replacement/refreshment strategies. A

[2] This facet is currently not available waiting for the Biofacets redesign to complete.

number of cache management schemes have been proposed and currently deployed by search engines such as Google, including [63-70]. These strategies mainly differ in terms of what data to cache and the data refreshing/replacement strategy. Biofacets strategy is mainly dictated by the first criteria as it deals with different types of data in terms of formats and levels of processing. The aim is to balance between the time necessary for internal processing, and the space available for data storage.

The solution we designed relies on storing both the internal representation (summary and extended) of the record and the URL. While the record URL is essential to retrieve the data, the argument on whether or not to store the record information locally is still in early stage. The experiments run on a limited data set clearly show the performance gain that the approach provides when compared to "no caching" policy. These results are supported by an efficient database design and heavy indexing support. However, more experiments need to be performed to assess the system scalability with the increased number of users and queries in order to determine a tradeoff between a satisfactory query response time and a manageable database. More studies and experimental support are needed to assess the adequacy of the proposed cache based on LRU (Least Resource Used) update strategy [71], especially as the system get deployed by the research community and the cache size limit is experienced in real time. Similarly, while the strategy of querying all supported databases seems to be appealing, especially that we provide the results to the users as soon as they are received by Biofacets, it remains to be tested to assess its impact on the system resources (see section 4).

4. Discussion

The Biofacets prototype demonstrates that the faceted search of biological databases is feasible. Such a tool should be advantageous to researchers. On the one hand, it provides results from many biological databases in one standard format, obviating the need for researchers to learn the varied interfaces of several biological database providers. On the other hand, Biofacets provides links back to the original data in the source databases if the researchers need to view these data. Biofacets is only a prototype and needs several enhancements.

As mentioned earlier, the current facets and sub-facets were manually identified. This process of finding facets could be semi-automated. We are currently investigating the use of clustering techniques to generate the faceted scheme. The initial results we obtained suggest that the fully automated faceted generation process needs knowledge expertise to guide the clustering process. This thread of research will be the part of the future research on Biofacets.

An additional enhancement would be to allow researchers to establish their own faceted scheme and then apply this scheme to the data. This may require the use of different technologies than are currently used in Biofacets.

Biofacets currently supports only a small number of biological databases. Many more databases need to be added to its repertoire.

As mentioned earlier, manual generation and maintenance of XSLT files and wrappers (to support HTML based databases) is not effective and will not scale to the numbers of biological databases available. These tasks need to be semi-automated and that work is already underway. In the context of the latter type of databases we are involved in a

research collaboration that is interested in using active learning [72] to propose a new scalable semi-automated approach to generate wrappers.
Finally, for Biofacets to be a truly usable tool, it needs to be accepted by the researchers who will be using it. Plans are being developed to allow various groups of potential users of Biofacets to experiment with Biofacets and provide their feedback. This feedback will be evaluated and incorporated into Biofacets as is feasible.

5. Acknowledgment

We would like to thank Myron Snelson, school of Informatics, IUPUI, for his insightful suggestions and help in proofreading the document.
This project was supported in part by NSF CAREERDBI-DBI-0133946 and NSF DBI-0110854.

6. References

[1] Hernandez, T. and S. Kambhampati, *Integration of biological sources: current systems and challenges ahead.* SIGMOD Rec., 2004. 33(3): p. 51-60.
[2] Goble, C. and R. Stevens, *State of the nation in data integration for bioinformatics.* Journal of Biomedical Informatics, 2008. 41(5): p. 687-93.
[3] Galperin, M.Y. and G.R. Cochrane, *The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.* Nucleic Acids Research, 2010. 39(Database): p. D1-D6-D1-D6.
[4] Vivano, F., et al., *Proteomic Biomarkers of Atherosclerosis.* Biomarker Insights, 2008. 3.
[5] Yueh, J. *Asbestos Litigation: Replacement Parts Doctrine Update*. 2011; Available from: http://pooleshaffery.wordpress.com/2011/02/21/asbestos-litigation-replacement-parts-doctrine-update/.
[6] W.R. Grace and executives charged with fraud, obstruction of justice, and endangering libby, montana community 2005; Available from: http://www.justice.gov/opa/pr/2005/February/05_enrd_048.htm.
[7] C. N. Wathen, R.M.H., *An examination of the health information seeking experiences of women in rural Ontario, Canada.* Information Research, 11(4) paper 267, 2006.
[8] S. Davidson, J.C., B. Brunk, J. Schug, V. Tannen, C. Overton and C. Stoeckert. *K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources*. in *IBM Systems Journal, 40(2), 512-531*. 2001.
[9] Stevens, R., P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N.W. Paton, C.A. Goble, and A. Brass, *TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources.* Bioinformatics. 16(2):184-186, 2000.
[10] Haas, L., Schwarz, P., Kodali, P., Kotlar, E., Rice, J, Swope, W., *DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources. IBM Systems Journal*, 40(2), 489-511, 2001.
[11] Zdobnov, E.M., R. Lopez, R. Apweiler and T. Etzold, *The EBI SRS server-recent developments. Bioinformatics*. 18(2): 368-373, 2003.
[12] Ben Miled, Z.L., N., Baumgartner, M., Liu, Y., *A Decentralized Approach to the Integration of Life Science Web Databases. Informatica*. 27(1). 2003.

[13] Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology. Nucleic Acids Research*. 31(1):28-33., 2003.

[14] Diaz, J.A., et al., *Patients' Use of the Internet for Medical Information.* Journal of General Internal Medicine, 2002. 17(3): p. 180–185-180–185.

[15] Christina, C., *Health information-seeking among Latino newcomers: an exploratory study.* Information Research, 10(2) paper 224, 2004.

[16] Wang, K., Tarczy-Hornoch, P, Shaker, R, Mork, P, Brinkley, J. *BioMediator Data Integration: Beyond Genomics to Neuroscience Data*. in *AMIA Fall 2005 Symposium Proceedings*. 2005.

[17] Hull, D., et al., *Taverna: a tool for building and running workflows of services.* Nucleic Acids Research, 2006. 34(Web Server issue): p. W729-732-W729-732.

[18] Girvan, M. and M.E.J. Newman, *Community structure in social and biological networks.* Proceedings of the National Academy of Sciences of the United States of America, 2002. 99(12): p. 7821-7826.

[19] Salwinski L, M.C., Smith AJ, Pettit FK, Bowie JU, Eisenberg D, *The Database of Interacting Proteins: 2004 update.* NAR 32 Database issue:D449-51, 2004.

[20] Sacco, G.M. *Research Results in Dynamic Taxonomy and Faceted Search Systems*. 2007; Available from: http://www2.computer.org/portal/web/csdl/doi/10.1109/DEXA.2007.75.

[21] Zhao, J., et al., *OpenFlyData: The Way to Go for Biological Data Integration*, in *Data Integration in the Life Sciences*. 2009. p. 47-54.

[22] Halevy, A., M. Franklin, and D. Maier. *Principles of dataspace systems*. in *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2006. Chicago, IL, USA: ACM.

[23] Jeffery, S.R., M.J. Franklin, and A.Y. Halevy. *Pay-as-you-go user feedback for dataspace systems*. in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008. Vancouver, Canada: ACM.

[24] Halevy, A.Y., *Answering queries using views: A survey.* VLDB, 2001. 10(4): p. 270-294.

[25] Suchanek, F.M., M. Vojnovic, and D. Gunawardena. *Social tags: meaning and suggestions*. in *Proceeding of the 17th ACM conference on Information and knowledge management*. 2008. Napa Valley, California, USA: ACM.

[26] Sacco, G.M. and Y. Tzitzikas, *Dynamic Taxonomies and Faceted Search Theory, Practice, and Experience*, ed. Springer. 2009.

[27] D. R. Cutting, D.R.K., J. 0. Pedersen, and J. W. Tukey, *Scatter/Gather: A cluster-based approach to browsing large document collections*, in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'92*. 1992. p. pp. 318-329.

[28] Meila, M. and D. Heckerman, *An experimental comparison of several clustering and initialization methods.* Machine Learning, vol. 42, no. 1/2, 2001: p. 9-29.

[29] H.-J. Zeng, Q.-C.H., Z. Chen, W.-Y. Ma, and J. Ma. *Learning to cluster web search results*. in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004*. 2004.

[30] Hearst, M.A. and J.O. Pedersen. *Reexamining the cluster hypothesis: scatter/gather on retrieval results*. in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. 1996: ACM.

[31] Zhao, Y., G. Karypis, and U. Fayyad, *Hierarchical Clustering Algorithms for Document Datasets.* Data Mining and Knowledge Discovery, 2005. 10(2): p. 141-168.

[32] Sanderson, M. and B. Croft. *Deriving concept hierarchies from text*. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999: ACM.

[33] W. Dakka and P.G. Ipeirotis, *Automatic discovery of useful facet terms*, in *SIGIR Faceted Search Workshop*. 2006.

[34] Emilia Stoica, Marti A. Hearst, and M. Richardson. *Automating creation of hierarchical faceted metadata structures*. in *NAACL-HLT 2007*. 2007. Rochester, NY.

[35] Anick, P., and Tipirneni, S. , *Method and apparatus for automatic construction of faceted terminological feedback for document retrieval*. 2003.

[36] Arentz, W.A. and A. Øhrn, *Multidimensional visualization and navigation in search results*, in *Proc. 8th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES'2004)*. 2004. p. 620--627.

[37] Hearst., M.A., *Design recommendations for hierarchical faceted search interfaces*, in *Proc. SIGIR 2006 Workshop on Faceted Search*. 2006. p. 26--30.

[38] Krellenstein, M.F., *Method and apparatus for searching a database of records*. 1999.

[39] Shneiderman, B., et al. *Visualizing digital library search results with categorical and hierarchical axes*. in *Proceedings of the fifth ACM conference on Digital libraries*. 2000.

[40] Meredith, D.N. and J.H. Pieper. *Beta: Better extraction through aggregation*. in *SIGIR'2006 Workshop on Faceted Search*. 2006.

[41] Wu, H., M. Zubair, and K. Maly. *Collaborative classification of growing collections with evolving facets*. in *Proceedings of the eighteenth conference on Hypertext and hypermedia*. 2007. Manchester, UK: ACM.

[42] Ben-Yitzhak, O., et al. *Beyond basic faceted search.* in *Proceedings of the international conference on Web search and web data mining*. 2008: ACM.

[43] Mahoui, M., Ben Miled, Z., Godse, A., Kulkarni, H., Li, N., *BioFacets: Faceted Classification for Biological Information*, in *Proc. of the 3rd International Workshop on Data Integration in the Life Sciences*. 2006. p. 104-113.

[44] Mahoui, M., Ben Miled, Z., Godse, A., Kulkarni, H., Li, N., *BioFacets: Integrating Biological Databases using Facetted Classification*, in *Proc. of the 15th International Conference on Software Engineering & Data Engineering,*. 2006. p. 205-210.

[45] Mahoui, M., Cheemalavagupalli, K.N., Padmanabhan, A.S, *Querying and Dynamically Classifying Biological Data: on the Issue of Performance*. 2007, School of Informatics - internal report.

[46] Kushmerick, *Wrapper induction: Efficiency and expressiveness.* Artificial Intelligence Journal, 2000.

[47] *NCBI utilities*. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils.

[48] *EBI*. Available from: http://www.ebi.ac.uk/.

[49] *XSLT Transformations*. Available from: http://www.w3.org/TR/xslt.

[50] *NewT*. Available from: http://www.ebi.ac.uk/newt/.
[51] Mahoui, M., et al. *BioFacets: Faceted Classification for Biological Information*. in *Scientific and Statistical Database Management, 2006. 18th International Conference on*. 2006.
[52] Hearst, M., *Design recommendations for hierarchical faceted search interfaces*, in *ACM SIGIR 2006 Workshop on Faceted Search*. 2006.
[53] Stevens, R., Goble, C.A., and Bechhofer, S., *Ontology-based Knowledge Representation for Bioinformatics. Briefings in Bioinformatics.* Briefings in Bioinformatics, 2000. 1(4): p. 398-416.
[54] *GO ontology*. Available from: http://www.geneontology.org/.
[55] Karp, P.D., Riley, M., Saeir,M., Paulsen, I.T., Paley, S., and Pellegrini, A., *Toole. The Ecocyc Database* Nucleic Acids Research, 30(1):56(8), 2002.
[56] Ben Miled, Z., N. Li, G. Kellett, B. Sipes and O. Bukhres., *Complex Life Science Multidatabase Queries.* Proceedings of the IEEE, 2002. 90(11).
[57] Wroe, C., R. Stevens, C. Goble, A. Boberts, M. Greenwood, *A Suite of DAM+OIL ontologies to Describe Bioinformatics Web Services and Data.* International Journal of Cooperative Information Systems, 2003. 12(2).
[58] Ben Miled, Z., M. Mahoui , N. Gao, L. Lu, J. Chen and Y. He., *A Service Discovery Approach in Support of Web Service Integration.* Proceedings of the 5th IEEE Symposium on Bioinformatics and Bioengineering, 2004.
[59] *NCBI Blast*. Available from: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi.
[60] *DBcat*. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102454.
[61] *UniProt*. Available from: http://www.uniprot.org/.
[62] McKenna, R.W., *Multifaceted approach to the diagnosis and classification of acute leukemias.* Clinical Chemistry, 2000. 46(8 Pt 2): p. 1252-1259.
[63] Arasu, A., J. Cho, H. Garcia-Molina, A. Paepcke, S. Raghavan, *Searching the Web.* ACM Transactions on Internet Technology, 2001.
[64] Markatos, E.P., *On Caching Search Engine Query Results.* Proc. of the 5th International Web Caching and Content Delivery Workshop, 2000.
[65] Silvestri, F., et al. , *A Hybrid Strategy for Caching Web Search Engine Results.* Proc. of the WWW conference (WWW 2003), 2003.
[66] Long., X., Suel, T., *Three Level Caching for Efficient Query Processing in Large Web Search Engines.* . Proc. of the WWW conference (WWW 2005), 2005.
[67] Jiang, S., Ding, X., Chen, F, *DULO: An Effective Buffer Cache Management Scheme to Exploit Both Temporal and Spatial Locality. Proc. of the 4th USENIX Conference on File and Storage Technologies (FAST'05).* 2005: p. pp. 14-16.
[68] Lempel, R.M., S, *Competitive caching of query results in search engines.* Theoretical Computer Science,323(2-3), , 2004: p. pp. 253 - 271.
[69] Feder , T., Motwani, R., Panigrahy, R. Zhu, A., *Web caching with request reordering.* Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms, 2002: p. pp.104-105.
[70] Lempel, R., Moran, S. , *Predictive caching and prefetching of query results in search engines.* Proc. of the 12th international conference on World Wide Web, 2003.

[71] Silberschatz, A., Galvin, P.B., Gagne, G., *005. Operating System Concepts*. 2005: John Wiley and Sons.
[72] Muslea, I., S. Minton, and C.A. Knoblock, *Active Learning for Hierarchical Wrapper Induction*, in *National Conference on Artificial Intelligence - AAAI*. 1999.

Advanced Biomedical Engineering

Edited by Dr. Gaetano Gargiulo

ISBN 978-953-307-555-6
Hard cover, 280 pages
Publisher InTech
Published online 23, August, 2011
Published in print edition August, 2011

This book presents a collection of recent and extended academic works in selected topics of biomedical signal processing, bio-imaging and biomedical ethics and legislation. This wide range of topics provide a valuable update to researchers in the multidisciplinary area of biomedical engineering and an interesting introduction for engineers new to the area. The techniques covered include modelling, experimentation and discussion with the application areas ranging from acoustics to oncology, health education and cardiovascular disease.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mahoui M., Oklak M. and Perumal N. (2011). Multi-Faceted Search and Navigation of Biological Databases, Advanced Biomedical Engineering, Dr. Gaetano Gargiulo (Ed.), ISBN: 978-953-307-555-6, InTech, Available from: http://www.intechopen.com/books/advanced-biomedical-engineering/multi-faceted-search-and-navigation-of-biological-databases

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen