eHealth Beyond the Horizon – Get IT There
S.K. Andersen et al. (Eds.)
IOS Press, 2008

A Method for Indexing Biomedical Resources over the Internet

Guillermo DE LA CALLE [a], Miguel GARCIA-REMESAL [a] and Victor MAOJO [a]
[a] *Biomedical Informatics Group, Artificial Intelligence Lab., Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain*

Abstract. A large number of biomedical resources are publicly available over the Internet. This number grows every day. Biomedical researchers face the problem of locating, identifying and selecting the most appropriate resources according to their interests. Some resource indexes can be found in the Internet, but they only provide information and links related to resources created by the owner institution of each website. In this paper we propose a novel method for extracting information from the literature and create a Resourceome, i.e. an index of biomedical resources (databases, tools and services) in a semi-automatic way. In this approach we consider only the information provided by the abstracts of relevant papers in the area. Building a comprehensive resource index is the first step towards the development of new methodologies for the automatic or semi-automatic construction of complex biomedical workflows which allow combining several resources to obtain higher-level functionalities.

Keywords. Data Acquisition, Data Capture, Classification, Indexing

Introduction

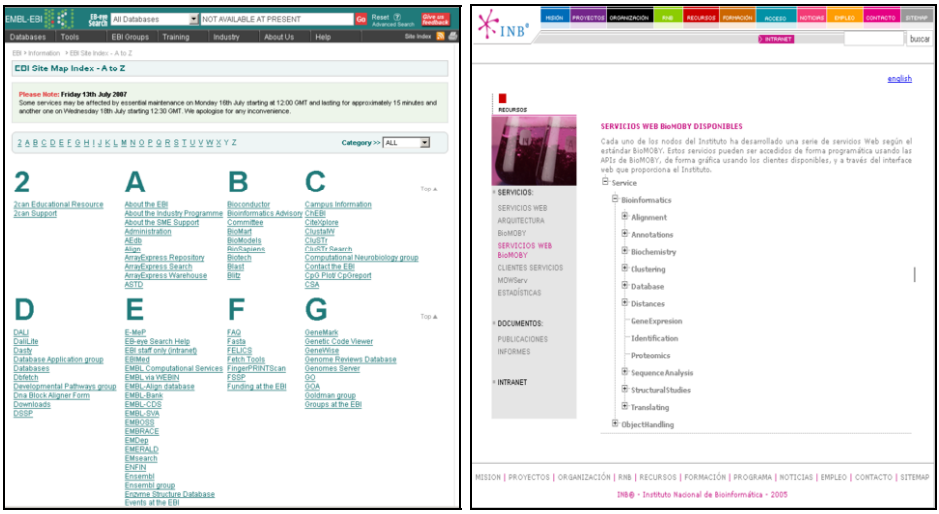
Over the last years, there has been a proliferation of biomedical resources, including databases, tools, and services. Many of these resources are publicly available over the Internet. They provide biomedical researchers with tools to facilitate different tasks, including, for instance, genetic sequence analysis and alignment, protein annotation or structural studies. As regards, there is a plethora of biomedical databases covering different areas including biomedical literature, genetic disorders, macromolecular structures or rare diseases. In this scenario, there is a need for novel tools to gather and organize all these resources. In most cases, the resources can be unknown to those researchers that may benefit from their use. This circumstance emphasizes the need for developing advanced tools to facilitate the localization and use of such resources.

In order to disseminate and provide a unified access to these resources, it is required to build an index of existing biomedical resources. Such an index should not only contain information regarding the location, inputs, outputs and service invocation procedures, but also a complete semantic description of their functionalities. Currently, there is not such an index readily available to the biomedical community.

1. Background

Numerous initiatives involving the development and availability of biomedical resource indexes can be found in the literature and on the Internet. For instance, the European Bioinformatics Institute (EBI) provides through his web-site a list of resources sorted alphabetically and grouped into three main categories: services, tools and databases. Some of them are provided directly by the EBI, and other are just links to the actual resources hosted by other institutions. Figure 1(a) shows a screenshot of the index published by the EBI. As shown in the figure, it is possible to search for a resource based on (i) the resource's name and (ii) type of resource, i.e. services, tools or databases. Unfortunately, the resources are not annotated with a semantic description of their functionality. Therefore, it is not possible to search for all the available resources needed for a concrete task – e.g. protein annotation or sequence alignment – in an efficient and intuitive manner. The only possibility is to manually inspect all the available resources to determine whether or not they meet the user needs. This is a time-consuming task that many biomedical researchers are reluctant to carry out.

Conversely, Figure 1(b) shows a more detailed index of resources – called BioMOBY – created by the Spanish National Institute of Bioinformatics (INB). In this case, all the resources are offered as web services and have been classified based on their functionality. For instance, as shown in the figure, they provide links to different bioinformatics resources, including alignment, annotations, clustering, databases, etc. When a user clicks on one of these links, she is forwarded to a web page with information about the service, such as name, type of resource, short description, inputs, outputs or location. The main drawback of this index is that it only provides access to



(a) (b)

Figure 1. Examples of resources indexes at the (a) European Bioinformatics Institute (EBI) and the (b) National Institute of Bioinformatics (INB) in Spain

resources owned by the INB. Valuable external resources from other institutions are not included in this index.

Other additional examples of resource indexes over the Internet are, for instance (i) the database collection mentioned by Galperin [1], and (ii) the list of resources available at [2]. The former reference [1] compiles a collection of publicly available molecular biology databases. This compilation is reviewed and published periodically every year. The latest version gathers a list of almost one thousand databases. The latter are classified into different categories according to their contents sorted in alphabetical order. For each database, the name, a short description of their contents and the link to the database is provided. The list of resources provided by (ii) is not only composed by databases but also tools, services and links to literature. More than two thousand resources have been classified into this list, depending on its functionality. In this paper, we present a novel semi-automatic method to create and maintain an index of biomedical resources. Figure 2 outlines the proposed method to create the Resourceome index.

2. Methods

A common problem for readily available indexes is that resources are annotated manually or they are not annotated at all. Creating or updating those indices by hand is a time-consuming task. Thus, it is a significant challenge to develop new methods and tools to facilitate the annotation and indexing of the resources. Another problem is the lack of standard domain models covering all required types of resources and tasks. All existing indexes use their own classification criteria or taxonomies that are neither complete nor do they use standard terminology to name their components.

Therefore, the first step would be the creation of a domain model or ontology not only covering the taxonomy of concepts but also all relevant relationships between the concepts. Domain knowledge provided by the relationships might contribute to a better understanding of the functionally provided by the resources and can be useful for resource classification and to enhance searches. According to Cannata et al. [3], ontologies are a suitable tool to annotate the resources and building the index. Firstly, they propose the need for creating an ontology including high-level concepts, attributes, and standard relationships between concepts. Next, biomedical researchers must be provided with a mechanism to maintain and extend the ontology with subconcepts to better describe their resources. Finally, the index built upon such ontology should be disseminated, enabling software engineers to create interfaces for searching and managing the resources.

First of all, it is necessary to obtain a list of the names of all the available resources that may be of interest for the biomedical community. Resources include databases, tools, and services. This list is created by an automated software agent that extracts the resource names from different online sources. Additional resource names can be added manually to the list by human annotators, if required. Once the list of resources has been obtained, it is necessary to extract detailed descriptions about their functionalities. This information can be usually found in abstracts of published papers or technical reports describing the resources. In many cases it is not necessary to examine the entire documents to extract this information, since abstracts often contain all the relevant information about the functionality provided by the resources. Once we obtain all the relevant abstracts, these are stored in a document repository. It is possible to store more

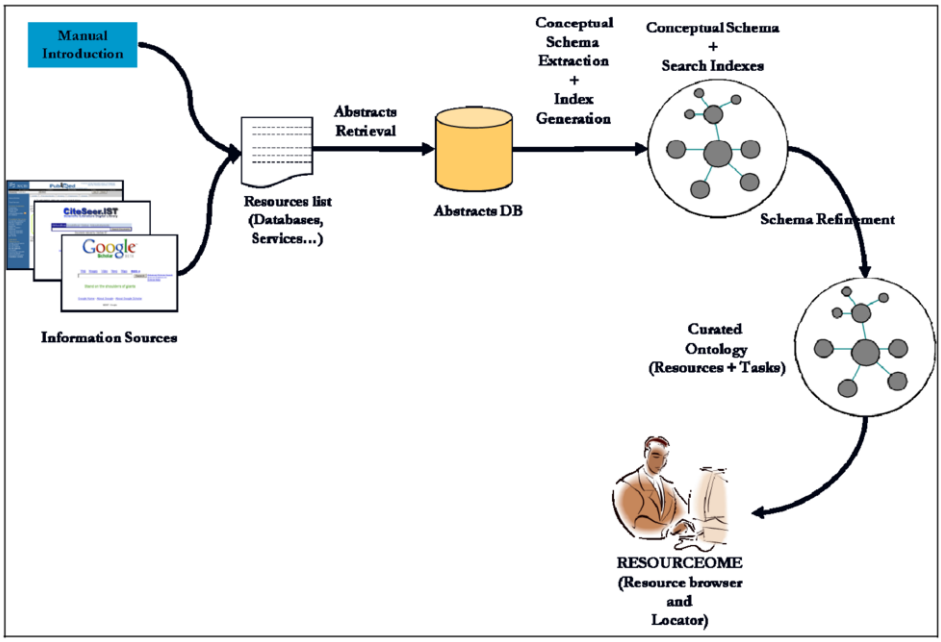


Figure 2. Overview of the "Resourceome" index creation method

than one abstract for a given resource in the database thus providing a more detailed description of the resource. Using the abstracts database, we automatically create a conceptual schema that describes the domain covered by the abstract database. This schema includes concepts – resources and tasks – as well as relationships between concepts. In this schema, relationships are of a single type – i.e. they denote which resources are useful for concrete tasks. The schema is extracted automatically from the selected abstracts using an original method developed by the authors and published elsewhere [4]. The method, based in different Artificial Intelligence (AI) and natural language techniques – including tokenizers, probabilistic part-of-speech taggers and transition networks – has been successfully used to bridge together schema-less, non-structured sources with relational databases. For further details on this issue see [4].

Our method works as follows. Each abstract is subdivided into phrases which are treated separately. Resources names, functionalities and all the relevant information is extracted by applying a special transition network built for this purpose. Besides the conceptual schema, it is necessary to create an index collection that relates the available resources to concepts in the conceptual schema. Using these indices, it is possible to browse the generated model to search for all the available tools to carry out a given task or databases containing a concrete type of data. Domain knowledge provided by the relationships can be exploited to enhance the searches.

Once the conceptual schema has been created, it is necessary to perform a refinement process to ensure the quality of the generated schema. The refinement task includes the removal of non-relevant concepts and relationships and naming concepts in the schema using a standardized terminology. To perform this task – which must be carried out by experts in the domain – it is possible to use widely accepted biomedical

vocabularies such as, for instance, the Unified Medical Language System (UMLS) [5], the Gene Ontology [6] or the Human Gene Nomenclature [7].

The refined model or ontology can be visualized by users using a graphical ontology viewer. We are currently developing an application programming interface (API) to provide access and search services for the Resourceome. Regarding the maintenance of the Resourceome, it is possible to incrementally update the index with new resources simply by adding new abstracts to the database. The system automatically creates the indices for the added resources. Next, the existing conceptual schema is automatically updated with concepts extracted from the new abstracts, not currently belonging to the conceptual model. Finally, experts can manually update the refined model – if required – to reflect the changes introduced by the addition of new resources to the Resourceome.

3. Results

In order to evaluate the model proposed in the previous section, we performed an experiment using a controlled input. We selected such kind of input since the AI and Natural Language Processing (NLP) techniques used in the development require some tuning to be adapted to the biomedical domain. This tuning is normally performed using controlled inputs to the system. We have designed this experiment targeted to tune and refine the model prior to execute more comprehensive tests with different data sources. The starting point was the list of resources offered by the EBI through its official website, previously mentioned. The list is organized into three categories (databases, tools and services). The names of fifty resources were randomly selected among the three categories and they were compiled into a candidate list. For each name of the list, the most relevant paper describing the resource was chosen from a public repository. For this experiment, we have used the ISI Web of Knowledge© [8] as the information source.

Once all the abstracts were manually retrieved, several AI and NLP techniques were applied to extract the relevant information contained in the abstracts. We have previously worked on a similar topic targeted to database integration and information retrieval [4]. For the present work, we have used the same approach for concept extraction and index generation. The target was focused in obtaining the name of the resources and their functionality. In the case of the databases, the detection of the functionality was substituted by the type of the data contained in the records. Additionally, extra information such as inputs, outputs, as well as relations with other resources was also searched. However, such information was only detected in few cases. Abstracts do not usually contain complete descriptions of the resources and such extra extraction is not always possible. Hence, we determined that an analysis of the full papers would be needed. Information extracted was automatically stored into a database specially built for this purpose in order to facilitate data analysis.

Some interesting results were discovered after analyzing the extracted data. First, the names of the fifty resources were properly extracted from the abstracts as well as their functionalities. A limited set of types or categories of resources has been established. In 94% of the cases, this category was correctly obtained from the functionality. Sometimes more than one category was found for a concrete resource. In such a case, both types were stored. These categories or concepts have been used to build a taxonomy for allowing the automatic classification of resources.

4. Conclusions and Future Research

The new model proposed is still in its early stages. Additional improvements and experimentation are needed. Results obtained in the preliminary experiments are promising and they suggest that the approach might be useful. Several information sources available on the Internet are currently being analyzed to be incorporated within the system in the next stages. Although some processes exposed before have been manually performed at this stage of the project, we are considering new approaches for automating the methods. We believe that the proposed method is an adequate starting point to organize the idea of a biomedical Resourceome.

Besides annotating and organizing the biomedical resources, the Resourceome might be a valuable tool to facilitate the construction of biomedical workflows. A workflow is a loopless forward chaining graph of available resources targeted to achieve a complex task. Workflows are usually created manually using tools such as Taverna [9], or Semantic Moby [10]. These tools provide workflow designers with a framework to create and execute scientific workflows in a graphical and intuitive manner. The main drawback of the manual workflow creation approach is that workflow designers need to have a good knowledge of the available resources as well as their required inputs and outputs. This approach can be suitable for simple or *ad-hoc* purposes, but it is not practical for more complex scenarios. The possible applications of creating automatically such workflows for the proposed Resourceome in different scientific areas is a great challenge for extending and applying the model presented.

Acknowledgments

The present work has been funded by the Ministry of Education, Spain (OntoMineBase project, reference TSI2006-13021-C02-01) and the E.C. (the INFOBIOMED NoE (FP6-2002-IST-507585) and the ACGT integrated project (FP6-2005-IST-026996)).

References

[1] M.Y. Galperin, *"The Molecular Biology Database Collection: 2007 update"*, Nucl. Acids Res. 35:D3-D4, 2006.
[2] Bioinformatics Links Directory.http://bioinformatics.ca/links_directory/ (last accessed Feb. 15th, 2008).
[3] N. Cannata, E. Merelli, R.B. Altman, "*Time to organize the bioinformatics resourceome"*, PLoS Comput Biol 1(7):e76, 2005.
[4] M. García-Remesal, V. Maojo, J. Crespo, H. Billhardt, "Logical Schema Acquisition from Text-Based Sources for Structured Biomedical Sources Integration". AMIA Annu Symp Proc. 2007;259-263, 2007.
[5] D.A.B. Lindberg, B.L. Humphreys, A.T. Mccray, "*The Unified Medical Language System"*, Methods of Information in Medicine 32(4):281-291, 1993.
[6] The Gene Ontology Consortium, "*Gene Ontology: tool for the unification of biology"*, Nature Genetics 25(1):25-29, 2000.
[7] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, H. Wain, "*The HUGO Gene Nomenclature Committee (HGNC)"*, Human Genetics 109(6):678-680, 2001.
[8] ISI Web of Knowledge. http://isiknowledge.com (last accessed Feb. 15th, 2008).
[9] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, et al, "*Taverna: a tool for the composition and enactment of bioinformatics workflows"*, Bioinformatics 20(17): 3045-3054 Nov 22, 2004.
[10] M.D. Wilkinson, D. Gessler, A. Farmer, L. Stein, "The BioMOBY Project Explores Open-Source, Simple, Extensible Protocols for Enabling Biological Database Interoperability". Proc Virt Conf Genom and Bioinf (3):17-27, 2003.