Towards Openness in Biomedical Informatics

Victor Maojo[1], Ana Jimenez-Castellanos[1], Diana de la Iglesia[1]

[1] Dept Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo S/N, 28660 Boadilla del Monte, Madrid, Spain.
{vmaojo, ajimenez, diglesia}@infomed.dia.fi.upm.es

Abstract. Over the last years, and particularly in the context of the COMBIOMED network, our biomedical informatics (BMI) group at the Universidad Politecnica de Madrid has carried out several approaches to address a fundamental issue: to facilitate open access and retrieval to BMI resources —including software, databases and services. In this regard, we have followed various directions: a) a text mining-based approach to automatically build a "resourceome", an inventory of open resources, b) methods for heterogeneous database integration —including clinical, -omics and nanoinformatics sources—; c) creating various services to provide access to different resources to African users and professionals, and d) an approach to facilitate access to open resources from research projects.

Keywords: Biomedical Informatics. Information retrieval. Web 2.0. Semantic Web.

1 Sharing

Information about Medical Informatics (MI) and Biomedical Informatics (BI) resources has dramatically grown over the past decade. A broad interest in these fields is leading professionals to produce new materials that can be shared and exchanged with the rest of the scientific community. To address this rapid growth, it is important to collect information and tools using automatic methods. In this regard, our group at the UPM has been working on a series of topics, in the context of the COMBIOMED network and the ACTION Grid Project, where various members of COMBIOMED also actively participated.

The Human Genome Project —and other -omics projects, as well— strengthened collaborative links among remote institutions that share and exchange software and data across remote organizations. In contrast, most clinical databases cannot be openly accessible due to privacy issues involving confidential patient information. In this context, there is already an extended amount of open-source software tools, created for tasks such as e-learning, in many disciplines. In biology, some examples are BioLogica (for genetics) and Dynamica (for kynematics). In medicine, there are currently proposals for a Medical Wikipedia and various public sources of medical images. Pubmed is the on-line, free access gateway to Medline, a comprehensive bibliographic reference for biomedical researchers and professionals. Medline was a

pay-per-service resource for various decades, until it became freely available for the biomedical community in the 1990's. Medline has had an enormous impact in biomedical research, education and practice.

Different Web technologies have been proposed for accessing and sharing remote heterogeneous information from open source tools. In a recent publication [1], members of our group at the UPM have proposed a new method to deal with this challenge. We reviewed several indexes of bioinformatics resources, currently available over the Internet. For instance, BioPortal [2], a web-based repository of biomedical ontology resources, developed by members of the US National Center for Biomedical Ontology. This application supports collaborative development of biomedical ontologies. BioPortal includes, among others, the Open Biomedical Resources (OBR) service for annotating and indexing biomedical resources. Resources are annotated by using a domain ontology. Other examples of such indexes include the Bioinformatics Links Directory (BLD) —a catalogue of links to bioinformatics resources, tools and databases classified into eleven major categories— where resources can be searched using keyword-based queries. —a catalogue of links to bioinformatics resources, tools and databases classified into eleven major categories— where resources can be searched using keyword-based queries. Resources are classified according to the type of service they provide —such as databases, tools and (web) services. The index includes both internal and external resources. A consortium of various US National Centers for Biomedical Computing has recently developed another index of bioinformatics resources called iTools [3]. A web-based interface enables researchers to locate the resources they need using advanced search and visual navigation tools.

Web-based repositories of bioinformatics resources have been built to facilitate their access to researchers in the area. Until now, these systems have been developed and updated manually. In this regard, our group proposed a new method, recently reported in a major scientific journal and conference [4], to automate this process. Informatics tools will then become available for the biomedical community and interoperable in actual research scenarios related to the VPH.

We describe below the fundamentals of this method.

BIRI (BioInformatics Resource Inventory) is a web application that allows users to search for bioinformatics resources (tools, frameworks, repositories, etc). Searches can be filtered by resource name, category or domain. Resources are classified according to a taxonomy of 9 domains and 28 categories. Domains represent the area of influence/application of resources—e.g. DNA, RNA or proteins—and categories denote the resource functionality or type—e.g. annotate, analyze or database. That taxonomy is based on other existing classifications such as the Bioinformatics Links Directory. A novel methodology has been developed to create the BIRI repository from the scientific literature based on Natural Language Processing (NLP) and Artificial Intelligence techniques. These methodologies allow retrieving, discovering and indexing resources automatically from manuscripts published in specialized journals in the bioinformatics domain. Extracting information from published papers guarantees that only relevant and peer-reviewed resources are indexed. Name, functionality and URL of the resources are directly extracted from the text (title and abstract). Additionally, resources are automatically annotated with one or several

categories and domains according to the BIRI taxonomy, depending on the textual description contained in the manuscript.

The methodology used to create the BIRI repository consists of five main phases:

1) Manuscript selection & surrogate generation.
2) Surrogate pre-processing.
3) Information extraction.
4) Resources classification.
5) Curation.

The BIRI approach presents several advantages over similar existing indexes: i) discovery and classification of resources is performed automatically, ii) the repository of resources can be updated by just feeding the system with new papers, iii) additional information sources might be used such as PubMed or Google Scholar, and iv) advanced search capabilities are provided through the web interface. Given the general methodology used in BIRI, a similar approach might be applied in other domains. Currently, some tasks, besides the automatic method, must be carried out, such as manuscript selection, taxonomy creation or final curation.

However, whereas sharing data and software tools is frequent in Bioinformatics, Medical Informatics is a discipline where there is an ongoing, long debate about medical Open Source Systems (OSS). One future realistic possibility is to have a pool of medical software systems which can be used on demand, and paid per use. Professionals can access these tools, use them and decide if they want to continue working with them. Such a scenario, proposed by Mandl and Kohane from Harvard, needs a platform and an infrastructure to become feasible [5]. This area of Open Source will surely become a hot topic in the coming years, particularly in the context of the Web 2.0.

2 Web 2.0 and 3.0

The past ten years, the idea of interactivity evolved from linking and clicking documents to creating and sharing. Thus, the Web 2.0 has been proposed to facilitate communication and simultaneous work between in different groups. Below is a summary of the differences between the Web 2.0 and its previous version:

Table 1. Differences between the Web 1.0 and Web 2.0

Web 1.0	Web 2.0
Application-based	Web-based
Isolated	Collaborative
Offline	Online
Licensed or purchased	Free
Single creator	Multiple collaborators

Although the use of the WWW is commonly related to searching for information, this new Web 2.0 infrastructure has enormous potential for developers and practitioners. Medical digital libraries, distributed medical records or Geographical Information Systems for medical issues —like Google maps used to graphically

represent the expansion of pandemias— are among the envisioned applications to be collaboratively developed and used by health professionals. While physicians were the initial targets and users of Web-based medical applications, patients are also demanding new applications to improve the quality of medical care. Using the Web, they aim to access second medical opinions, find personalized advice or contact their physicians or other patients directly. A new version of the WWW, called the Semantic Web —or Web 3.0— emphasizes the use of semantic-based technologies for organizing and structuring the Web by means of ontology-related technologies. Such a new approach facilitates tasks such as information storage, retrieval or mining. We have also reported various semantic-based research and technologies [6,7].

3 Importance of Medical Information Systems for developing countries

In such expanded context of the Web 2.0, we have carried out an analysis of activities related to BMI in Africa. For this work, collaboration with an expert from Egypt, Dr. Rada Hussein, has been fundamental. An analysis of the literature made by means of BIRI and related tools, combined with manual Medline and Google searches, has suggested enormous opportunities and challenges for transferring results from many previous EU research projects in BMI to African locations for improving medical practice and research. We have to remind that, in the ICT for Health area of the EC, there have been few contacts with African BMI professionals. Thus, there is a great room for improvement.

Fig. 1. Members of UPM teaching computer science to young students in Burundi

Global health has experienced significant developments, but efforts for cooperation with underdeveloped countries must increase. Countries like China and others have largely improved their health indicators, recently, compared to rich countries, where inequalities can still be widely found. In this context, institutions such as WHO —a collaborator of our group— have established priorities for improving global health over various decades. These priorities depend on accurate numbers and estimations

extracted from public health systems, which are still largely unknown in many African countries.

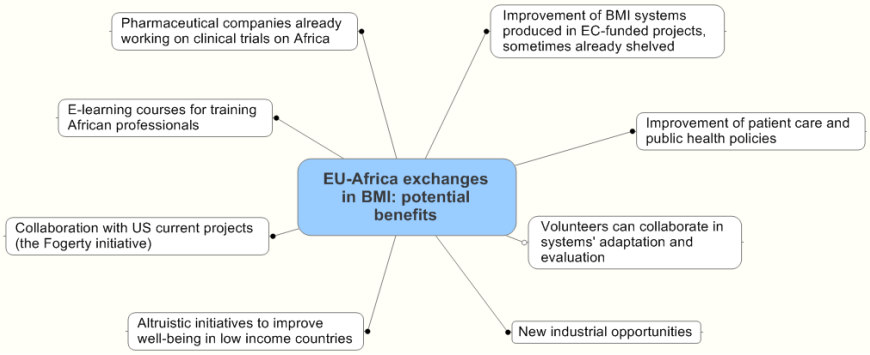

Fig. 2. Representation of examples of potential benefits from a transfer of knowledge and systems in BMI from Europe to Africa

4 Costs of medical technology: Information systems

Research on medical technology has had an enormous impact on medicine. In fact, it has been usually considered that one of the most significant issues defining modern medicine is the advance of medical technologies. Within these technologies, we will focus on medical information systems.

In 2002, a survey carried out at the Health Affairs journal among 225 medical internists did not consider the use of computers in medicine among the 30 most significant medical innovations of the last decades. Nevertheless, a few years later, another survey carried out over Internet by the British Medical Journal ranked the use of computers in medicine in the 10th place among the most significant advances in medicine since the journal was created in 1840.

Table 2. An extract of the results of the British Medical Journal survey in 2007

Sanitation	15.8
Antibiotics	14.5
Anesthesia	13.9
Vaccines	11.8
Discovery of DNA structure	8.8
Germ theory	7.4

Although these kinds of surveys should be carefully considered, this last result may indicate a significant shift in the consideration of the use of computers in medicine.

In fact, if we consider the time that medical professionals dedicate to information management-related tasks, it has been earlier observed —as soon as in 1966— that these rates are quite large, ranging from 95% for medical records professionals to 28% for laboratory workers. For physicians, these rates ranged from 30 to 36%. Since this survey was carried out in 1966, it can be hypothesized that time dedicated to these tasks may be quite higher now. Thus, information management is already assumed to be a fundamental component of modern medical practice.

Nevertheless, there are several aspects that can be considered, regarding medical information systems, which will surely have a positive effect in terms of cost control contention —a process which seems to have already began:

1. Medical Information systems development has reached a level of stability that allows the acceptance of many —current or "de facto"— standards by academics and industry —e.g., HL7, DICOM, UMLS, Web components, etc—, facilitating systems interoperability and components' reuse.
2. Costs and size of computer hardware have been continuously going down for various decades. Currently, a 20 euro pen-drive can store much more data than heavy storage juke-boxes 15 years ago, at lower than a 1% price and size. Plans to market personal computers in developing countries at a price below 100$ have been proposed for several years.
3. There is an increasing culture of developing open-source software systems and data sharing that have pervaded related disciplines such as genomics and bioinformatics, allowing many "-omics" projects to be completed before schedule. At the same time, there are many public databases offering free gene, protein and disease information to the scientific community. The completion of the Human Genome Project before schedule, due to collaborative efforts and data and software sharing among researchers all over the world, triggered the development of numerous publicly available databases containing gene, protein and disease as well as bioinformatics tools. This number is continuously increasing and it is now over 1300 public databases. Security, confidentiality and ownership rights have prevented to reach similar importance and numbers in medicine but an increasing culture of sharing data and software tools as wells as the development of techniques for issues such as reliable anonymization techniques for managing patient data could help to expand it. In this regard, an interesting trend is to store these databases using cloud computing techniques. For instance, Amazon is providing free storage for some publicly available scientific databases in the genomics area.

5 The explosion of medical information

The development of the World Wide Web has configured a new scenario where people exchange huge amounts of information in all domains. The success of the World Wide Web after 1990 —what could be called the Web 1.0, as mentioned above— caused an explosion of the amount of biomedical information available for practitioners and researchers, and also for patients and public in general. An

enormous amount of biomedical information, never seen before, has been available for health practice, policy-making and research.

In the past years a new approach has obtained an immediate success. Web services have been defined by the W3C, the WWW consortium, as "a software system designed to support interoperable machine-to-machine interaction over a network". Web services can be accessed over the Internet, and executed on a remote system. Using the appropriate standards, such as WSDL, SOAP and others, Web services have been developed for numerous applications, also in biomedicine. Many applications can be run and executed as services, without a strong computing expertise needed. Web services can be orchestrated by means of workflows, according to the needs of each user.

The development of the Semantic Web, the Intelligent Web, or Web 3.0, where documents and tools can be structured, shared and integrated through intelligent semantic techniques promises to expand the above ideas. By way of an example, a special interest group called 'Semantic Web for Health Care and Life Sciences Interest Group' was created by the W3C to analyse the impact of the Semantic Web in the biomedical domain.

As a summary, one of the fundamental goals for the forthcoming years will be to structure information to facilitate information search and retrieval. In this regard, regrettably, there are many results from past research in BMI that are difficult to find —even those which are in the open community. Thus, we have already proposed a strategy to facilitate how to access such resources, as presented below.

6 A proposal for making open results from biomedical research projects easy to find and access

Wald has addressed scientific openness in a recent Science article [8], including data and methods used for research. Advances in software tools for bioinformatics search helps [3], but, just becoming aware of open results of research projects funded by public agencies—e.g., databases, software, papers, e-books— and finding them efficiently still proves harder than it should.

In the course of producing an advanced, automatically generated on-line inventory of bioinformatics resources [1], we analyzed results from research projects publicly-funded by the European Commission, Spanish agencies and the NIH. We discovered that finding the complete set of available information reported to have been generated by the projects could prove quite elusive. Non-peer-reviewed summary reports were commonplace, but specifics of electronic resources with Web locations were frequently not, even when researchers mentioned their existence as being openly available [9].

To enable searches with sophisticated text mining, publicly-funded projects should provide a *minimum information set* including titles, authors, funding agency, annotations with concepts from ontologies or controlled vocabularies that characterize the functionalities of the resources, papers reporting significant findings using these

resources —peer-reviewed quality indicators— and their Uniform Resource Identifiers (URIs).

Earlier suggestions for structuring abstracts of papers [10] resulted in an experiment with disappointingly limited success [11]. However, to provide basic information resources from projects already on the web ought to be more straightforward. Requiring a minimum information set like the one we propose to be available online under clearly specified standards might help bring about more comprehensive open access which would promote wider reuse of resources and avoid duplication in scientific projects, worldwide. Agencies are increasingly requiring that papers reporting research funded by them become publicly available. Our proposal is that they require that other products of research —like open electronic resources that back-up a paper's results— should be made equally easily accessible. Similarly, the use of text mining techniques can avoid duplication and plagiarism in proposals, as we have proposed previously in a communication to the Nature journal [12].

References

1. De La Calle G, García-Remesal M, Chiesa S, De La Iglesia D, Maojo V : BIRI: A New Approach for Automatically Discovering and Indexing Available Public Bioinformatics resources from the Literature. BMC Bioinformatics , 10, 320 (2009)
2. Musen M, Shah N, Noy N, Dai B, Dorf M, Griffith N, Buntrock JD, Jonquet C, Montegut MJ, Rubin DL: BioPortal: Ontologies and Data Resources with the Click of a Mouse. AMIA Annual Symposium Proceedings 2008. 1223-1224. (2008)
3. Dinov ID, Rubin D, Lorensen W, et al. iTools: a Framework for Classification, Categorization and Integration of Computational Biology Resources. PLoS ONE, 3(5):e2265 (2008)
4. De la Calle G, García-Remesal M, Maojo V: A Method for Indexing Biomedical Resources over the Internet, Stud Health Technol Inform, 136,163-168 (2007)
5. Mandl KD, Kohane IS.: No small change for the health information economy. N Engl J Med. 360(13):1278-81 (2009)
6. Alonso-Calvo R, Maojo V, Billhardt H, Martin-Sanchez F, García-Remesal M, Pérez-Rey D.: An agent- and ontology-based system for integrating public gene, protein, and disease databases. J Biomed Inform.40(1):17-29 (2007)
7. Pérez-Rey D, Maojo V, García-Remesal M, Alonso-Calvo R, Billhardt H, Martin-Sánchez F, Sousa A.: ONTOFUSION: ontology-based integration of genomic and clinical databases. Comput Biol Med. 36(7-8):712-30. (2006)
8. Wald C.:Scientists Embrace Openness. Science Issues and Perspectives. Science. (2010)
9. Maojo V, Garcia-Remesal M, Crespo J, de la Calle G, de la Iglesia D, Kulikowski C.: Open results from biomedical research projects: where are they? ScienceCareers (a section of the Science journal). (2010)
10. Gerstein M, Seringhaus M, Fields S. Structured digital abstract makes text mining easy. Nature. 447(7141):142. (2007)
11. Lok C. Literature mining: Speed reading. Nature.463(7280):416-8. (2010)
12. Maojo V, García-Remesal M, Crespo J. Detectors could spot plagiarism in research proposals. Nature. 456(7218):30. (2008)