Contributor names and short CVs:

Francisco M. Couto (Associate Professor) obtained a PhD (2006) in Bioinformatics from the University of Lisbon. He graduated (2000) and has a master (2001) in Informatics and Computer Engineering from the IST-UTL. He was on the faculty at IST-UTL from 1998 to 2001, and has been since 2001 a professor at FFCUL and a member of LASIGE. Francisco M. Couto is currently the coordinator of the master in Bioinformatics and Computational Biology, and a member of LASIGE coordinating the XLDB research group and the Biomedical Informatics research line. He was an invited researcher at EBI, AFMB-CNRS and BioAlma during his PhD studies. He received the Young Engineer Innovation Prize 2004 from the Portuguese Engineers Guild. His current research focus is on scientific data and knowledge management, including text and data mining, semantic web, data integration and sharing, and biomedical informatics. He published 30 articles in scientific journals and his work has over 1350 citations since 2010. He participated in the EPIWORK project, suppported by FP7, coordinated the participation of LASIGE in the Virtual Physiological Human network of excellence, and is currently working on the project BiobankCloud: Scalable, Secure Storage of Biobank Data, supported by FP7.

Web page: http://webpages.fc.ul.pt/~fjcouto/

Type of the presentation proposed:

research contribution

Title of the presentation:

METARATE - incentivize data integration and sharing by rating and rewarding metadata annotation

Summary of the presentation:

This presentation will describe a disruptive approach that promotes and intensifies raw data sharing and integration by simple rewarding and recognizing metadata sharing and integration on the semantic web using ontologies. The approach measures the knowledge rating of a dataset according to the specificity and distinctiveness of its mappings to ontology concepts. These measures will calculate the IC and the conceptual similarity of those ontology mappings to other existing datasets. The knowledge ratings will then be used as the basis of a novel reward and recognition mechanism that will rely on a virtual currency, dubbed KnowledgeCoin (KC).

Extended abstract of the presentation

Promoting data integration and sharing is essential to avoid the creation of silos that store raw data that cannot be reused by others, or even by the owners themselves. For example, the current lack of incentive to share and preserve data is sometimes so problematic, that are even cases of authors that cannot recover the data associated with their own published works. However, the problem is how to obtain a proactive involvement of the research community in data integration and sharing. In 2009, Tim Berners-Lee gave a TED talk[1], where he said: "*you have no idea the number of excuses people come up with to hang onto their data and not give it to you, even though you've paid for it as a taxpayer*". Public funding agencies and journals may enforce the data-sharing policies, but the adherence to them is most of the times inconsistent and scarce. Besides all the technological advances that we may deliver to make data integration and sharing tasks easier, researchers need to be motivated to do it correctly. For example, due to the Galileo's strong commitment to the advance of Science, he integrated the direct results of his observations of Jupiter with careful and clear descriptions of how they were performed, which he shared in Sidereus Nuncius. These descriptions enabled other researchers not only to be aware of Galileo's findings but also to understand, analyse and replicate his methodology. Thus the commitment of the research community to data integration and sharing is currently a major concern, and this explains why BMSRIs have recently included in their definition of the principles of data management and sharing the following challenge: "*to encourage data sharing, systematic reward and recognition mechanisms are necessary*". They suggest studying not only measurements of citation impact, but also highlighting the importance to investigate other mechanisms as well. Systematic reward and recognition mechanisms should motivate the researchers in a way that they become strongly committed in sharing data, so others can easily understand and reuse it. By doing so, we encourage the research community to improve previous results by replicating the experiments and testing new solutions. However, before developing a reward and recognition mechanism we must formally define: i) what needs to be rewarded and recognized; ii) and measure its value in quantitative and objective way.

Proper data integration and sharing is more than storing the datasets in a public repository, it requires the data to be organized, characterized and updated continuously, so others can find it and reuse it effectively. In an interview to Nature, Steven Wiley[2] emphasized that sharing data "*is time-consuming to do properly, the reward systems aren't there and neither is the stick.*" Sharing data without being connected to other external resources hampers its efficient retrieval and analysis, and also its expansion and update. For example, a properly linked dataset will not become outdated so easily and therefore it will tend to maintain its initial value. Like the careful and clear descriptions provided by Galileo, semantic characterizations in the form of metadata must also be present so others can easily find the raw data and understand how it can be retrieved and explored. Metadata is a machine-readable description of the contents of a resource made through linking the resource to the concepts that describe it. E.g. a dataset links to the concept of "influenza" because it contains data concerning that disease. However, if we really are to fully understand such diverse and large collections of raw data being produced, their metadata need to be integrated in a non-ambiguous and computational amenable

[1] http://www.ted.com/talks/tim_berners_lee_on_the_next_web
[2] http://www.nature.com/news/2011/110914/full/news.2011.536.html

way. The complex process of enriching a resource with metadata by means of semantically defined properties pointing to other resources often requires human input and domain expertise. Thus, METARATE approach assumes that by rewarding and recognizing metadata sharing and integration on the semantic web using standard and controlled vocabularies, we are promoting and intensifying scientific collaboration and progress. However, we need to define the value of metadata in terms of knowledge it provides about a given dataset. Semantic interoperability is a key requirement in the realization of the semantic web and it is mainly achieved through mappings to resources that reliably represent the abstraction of real-world objects and their interactions. Metadata can then be considered as a set of links where all the links are equal, but some links are more equal than others (adaption of George Orwell's quote). Thus, METARATE aims at measuring the knowledge rating of any given dataset through its mappings to concepts specified in an ontology, which can be viewed as a collection of concepts and the relationships between them. These relationships provide concepts with a machine-readable meaning that can be explored for information retrieval, pattern recognition, knowledge discovery or any other computational analysis. Thus, the main goal of METARATE is to demonstrate that the metadata integration and sharing value of a dataset, dubbed as knowledge rating, is proportional to the specificity and distinctiveness of its mappings to ontology concepts in relation to all the others datasets.

The specificity of a set of ontology concepts can be defined by the information content (IC) of each concept. For example, intuitively the concept dog is more specific than the concept animal. This can be explained because the concept animal can refer to many distinct ideas, and, as such, carries a small amount of information content when compared to the concept dog, which has a more informative definition. The distinctiveness of a set of ontology concepts can be defined by its conceptual similarity to all the others sets of ontology concepts, i.e. a distinctiveness of a dataset is high if there are no other semantically similar datasets available. Conceptual similarity explores ontologies and the relationships they contain to compare their concepts and, therefore, the entities they represent. Conceptual similarity enables us to identify that arm and leg are more similar than arm and head, because an arm is a limb and a leg is also a limb. Likewise, because an airplane contains wings, the two concepts are more related to each other than wings is to boat.

METARATE project will undertake two steps: i) assess the knowledge ratings against a gold-standard; ii) and integrate the knowledge ratings in a reward and recognition mechanism. The gold-standard creation will be closely monitored by domain-experts to ensure its high-quality. Effective knowledge ratings will have to achieve a significant correlation with the curated ratings provided by the gold-standard. The reward and recognition mechanism will rely on the implementation of a new virtual currency, dubbed KnowledgeCoin (KC), that will be specifically designed to promote and intensify the usage of semantic web technologies for scientific data integration and sharing. The idea is that every time a scientific article is published, KCs are distributed according to the knowledge rating of the datasets supporting that article. Note that KCs will by no means be a new kind of money and the design of KC transactions will focus on the exchange of scientific data and knowledge. The goal is to achieve a high number of transactions, which means an intensification of data integration and sharing within the research community.