Editorial

Zoé Lacroix

Scientific Data Management Laboratory,
Department of Electrical Engineering,
Arizona State University, Tempe AZ 85281-5706, USA

and

Pharmaceutical Genomics Division,
Translational Genomics Research Institute,
13400 E. Shea Blvd, Scottsdale, AZ 85259, USA
E-mail: zoe.lacroix@asu.edu

Biographical notes: Zoé Lacroix received her PhD in Computer Science in 1996 from the University of Paris XI – Orsay (France) and worked at the French Institut National de la Recherche en Informatique et Automatique (INRIA), the Institute for Research in Cognitive Science (IRCS) at the University of Pennsylvania (USA), and two biotech companies Gene Logic and at SurroMed. She currently holds a joint appointment as an Associate Professor of Research at Arizona State University and an Associate Investigator at the Pharmaceutical Genomics Division of the Translational Genomics Research Institute (TGen).

1 Introduction

The First International Workshop on Resource Discovery (RED 2008) brought together researchers, developers, and practitioners to discuss research issues and experience in developing and deploying concepts, applications, and solutions addressing various issues related to resource discovery. Five research papers were peer-reviewed and selected for presentation at the workshop. Although all papers presented original solutions to support resource discovery that can be applied to any context, because of the dynamic research and development effort toward supporting resource discovery for the life sciences (Lacroix et al., 2008a, 2008b), many papers used evaluation scenarios from the life sciences and bioinformatics. The four papers included in this volume are extended versions of papers presented at the workshop. Among them, three papers address issues related to resource publication and classification while the fourth one focuses on the problem of querying and composing resources discovered on the web. *Facilitating Discovery on the Private Web using Dataset Digests*, by Peter Mork et al. addresses the problem of visibility of a data source. Their approach uses statistical metadata that summarise the data source so that potential users may evaluate whether the resource contains the data they need. In *BioRegistry: Automatic Extraction of Metadata for Biological Database Retrieval and Discovery*, Marie-Dominique Devignes et al. use conceptual clustering methods to index and classify biological databases. The problem of indexing resources with a domain ontology is addressed in *Resource Descriptions, Ontology, and Resource Discovery*. Finally, Sergio Mergen et al. propose a re-writing mechanism that translates an input query in the terms of resources available on the Web in *Querying Structured Information Sources on the Web*. Two papers invited from IIWAS discuss the problems specific to Web service discovery and matching. In *P2P-SDSD: On-the-fly Service-based Collaboration in Distributed Systems*, Devis Bianchini et al. discuss semantic driven matching over a P2P network whereas Ioan Salomie et al. focus on planning mechanisms in *Web Service Composition Using Fluent Calculus*. The last paper authored by Abdelkader Hameurlain, IIWAS keynote speaker, and two collaborators is a survey on resource discovery in grid architectures. In addition to this exceptional selection, I am pleased to report on some of the discussions that took place during the workshop as they reflect how resource discovery captures a variety of problems and viewpoints uniquely represented at IIWAS and the reason why ontologies and metadata constitute the key components of the solutions.

2 Discussion

2.1 What is a resource?

A *resource* is something that is useful, that provides a service. Some of the workshop attendees see a resource as a node on a grid or a something that can be addressed on the Web: a resource has a URL. By extension, some even claim that anything behind a URL is a resource. A proposal was to limit resources to entities that can be identified with a URI on the Web. Finally, some see a resource as anything that can be identified, addressed, and with an interface that specifies what one needs to supply for the resource to provide the service.

Then the type of service that a resource should be able to provide was discussed. For some a resource corresponds to an information source such as a document (including text, audio, video), a data repository, a database management

system, a data cube. For others, it is operational such as an application, a tool, and data are resources once they provide a tool to access their content (e.g., a query form or a textual search engine). A link between data sources such as an index or a hyperlink can also be seen as a resource. For some, a resource can be memory, CPU, virtual resources, IO, disk IO. It was noted that to be useful, a resource should be shared, should be found or sensed, and the type of service provided by a resource must be clearly identified and map the expected usage task. Finally, for others a resource is not a static item but rather mobile, changing locations, like agents do.

2.2 Resource description

Resources are characterised by core information including a name, a description of the input and output (parameters or format), address, and various additional properties expressed as *metadata*. Different formats such as URI, RDF, XML, SOAP, WSDL, WADL, were discussed by the workshop attendees. All attendees agree that current formats are not expressible enough to capture all levels needed to support resource discovery. In other words, potential users of resources need to gain access to more information that currently expressed by formats. For a user who sees a resource as an agent the representation as a Web service is far from sufficient as it does not capture any mobility.

The use of ontologies is a promising approach to capture the semantics of resources. The input and output of the resource may be captured in terms of concepts and the resource itself may be capturing a conceptual relationship. Such metadata would capture what the resource does and how it does it.

Data sources pose specific problems not yet properly addressed by the current formats. In particular, the description of their content is critical to support resource discovery. Solutions include the publication of a data sample, a summary, or similar data digest, meaningful indices and other metadata, statistics including cardinality, and textual descriptions. Two papers presented at the workshop addressed the problem of data source description. The problem of using representations such as Web services to represent data sources was also discussed. Web service data format can only represent one access to a data source. Therefore a database can only be represented through a limited number of queries, thus specific views. Some of the workshop attendees felt that it did not matter because in resource discovery the database only had to indicate it existed and what it contained; others expressed concerns that the use of a database as resource discovery is often combined with resource composition and workflow systems as discussed in Section 2.3.

2.3 Metadata to support resource composition

Resource discovery is typically coupled with systems that aim at composing resources into pipelines (linear composition) or complex workflows (networks of resource calls) such as Oinn et al. (2006). In this case, resource discovery queries may be driven by various other motivations such as format and syntax (in order to map a resource output to the next resource input). Although resource access through keywords is a useful functionality it fails at expressing the variety of characteristics and queries against metadata that are critical to capture the specific aim of each resource, its operational mode (performance), and the resource selected will impact the overall workflow.

Metadata are data that describe a resource. Metadata include a wide range of information from attribution metadata, such as those attributes defined in the Dublin Core, to detailed policy metadata indicating who can access the resource under what conditions. *Semantic metadata* include the description of a resource with respect to the domain knowledge. *Syntactic metadata* provide the description of the resource interface. *Summary metadata* describe the actual contents of the resource. These metadata include free text summaries and statistical summaries of the instances (values) contained in the database. Summary metadata can be classified along several axes:

- textual vs. quantitative
- structured vs. unstructured
- manually generated vs. automatically generated.

By far the most common type of summary metadata is textual. Textual metadata allow an application developer or end- user to search for resources using keywords or phrases. The success of existing approaches seems to show that it is a familiar and intuitive operation, which works well when searching for reasonably well-defined concepts. Textual metadata are unstructured (i.e., free text) and manually curated. Alternatively, summary metadata can take the form of keywords drawn from a *controlled vocabulary*. A controlled vocabulary makes it easier to search for resources, assuming the vocabulary is sufficiently expressive and used consistently to annotate the resources. In most cases, textual metadata are generated manually, although there is some research to extract automatically keywords from a resource for its annotation.

Metadata management relies on the description of resources including the resource name, identification, and all additional information that may be relevant to locating, evaluating, and using the resource. A *resource identifier* is a sequence of characters that uniquely identifies a resource and is globally shared and understood over a network. A resource is analogous to a node on the Web. Identifiers are assigned to resources so that they can be uniquely identified on the Web. The ubiquitous Uniform Resource Locator (URL) is an example of a resource identifier, which uses the location, the local directory path and the local file name of the resource to locate it on the Web. Unique Resource Identifiers (URIs) include URLs that not only identify the resource but describe its primary access mechanism or network location, and Uniform Resource Names (URN) that identify a resource by name in a particular namespace.

2.4 Issues in resource discovery

Resource discovery is the process of identifying and locating existing resources that have a particular property. Machine-based resource discovery relies on crawling, clustering, and classifying resources discovered on the Web automatically. Resources are organised with respect to metadata that characterise their content (for data sources), their semantics (in terms of ontological classes and relationships), their characteristics (syntactical properties), their performance (with metrics and benchmarks), their quality (curation, reliability, trust), etc. Resource discovery systems allow the expression of queries to identify and locate resources that implement scientific tasks.

Several issues were discussed including resource publication, resource comparison, discovery interface, and discovery query languages. To be discovered a resource must be published either by making it publicly available in a format that can be identified by robots and other crawling tools or by registering it in public repositories such as Seekda.[1] Such repositories should classify resources and offer various discovery method either graphical or query-based. A query language cannot be limited to search resources by name. It should rather allow the identification of resources that best meet the user needs.

3 Conclusion

The first occurrence of a workshop devoted to Resource Discovery was a success and the exciting discussions that took place identified various areas of research that spam across various domain expertises from Web services to databases via Semantic Web and ontologies and middleware and agent-based approaches.

The Second International Workshop on Resource Discovery (RED 2009) was joint to the 35th International Conference on Very Large Data Bases (VLDB) on August 28, 2009, in Lyon, France. The proceedings of this second edition will be published in a volume of Lecture Notes in Computer Science by Springer in 2010.

Acknowledgements

I would like to thank the workshop attendees who contributed enthusiastically to the workshop discussion and whose thoughts are combined in this report and particular thanks to Marie-Dominique Devignes who acted as a scribe to record these discussions. Warm thanks to the members of the program committee for their time and valuable contribution to the workshop. Many thanks to Ismail Khalil Ibrahim who kindly and efficiently supported this event and the whole IIWAS committee for wonderfully organising it. The Translational Genomics Research Institute (TGen) is also thanked for sponsoring the workshop.

References

Lacroix, Z., Kothari, C.R., Mork, P., Rifaieh, R., Wilkinson, M., Cohen-Boulakia, S. and Freire, J. (2008b) 'Biological resource discovery (to appear)', in Liu, L. and Ozsu, M.T. (Eds.): *Encyclopedia of Database Systems*, Springer-Verlag, ISBN 978-0-387-35544-3,978-0-387-39940-9, pp.220–223, 2009.

Lacroix, Z., Kothari, C.R., Mork, P., Wilkinson, M. and Cohen-Boulakia, S. (2008a) 'Biological metadata management', in Liu, L. and Ozsu, M.T. (Eds.): *Encyclopedia of Database Systems*, Springer-Verlag, ISBN 978-0-387-35544-3,978-0-387-39940-9, pp.215–219, 2009.

Oinn, T.M., Greenwood, R.M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K., Goble, C.A., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P.W., Pocock, M.R., Senger, M., Stevens, R., Wipat, A. and Wroe, C. (2006) 'Taverna: lessons in creating a workflow environment for the life sciences', *Concurrency and Computation: Practice and Experience*, Vol. 18, No. 10, pp.1067–1100.

Note

[1]http://seekda.com/