Research Object Management: Opportunities and Challenges

Khalid Belhajjame
School of Computer Science
University of Manchester
Manchester, UK
khalidb@cs.man.ac.uk

David De Roure
Oxford e-Research Centre
University of Oxford
Oxford, UK
david.deroure@oerc.ox.ac.uk

Carole A. Goble
School of Computer Science
University of Manchester
Manchester, UK
Carole.Goble@manchester.ac.uk

ABSTRACT

While electronic papers have played and continue to play a primordial role in the dissemination of research results, researchers now recognize that papers are by no means sufficient to communicate and share research results. As a step in this direction, we present research objects as an abstraction for communicating, sharing and reusing research results. As well as the paper describing the contribution made by the scientist, a research object bundles information about the hypothesis the scientist investigated, the workflow implementing the experiment ran to assess the hypothesis, the data set used, the results obtained, and the conclusions drawn by the scientist, and identify a set of research problems that together aim to enable the management of research objects. We also underline the important role that end-users and automation techniques can play to enable scalable management of research objects.

INTRODUCTION

Research is increasingly digital. Most of research results are nowadays disseminated in the form of electronic papers through traditional communication channels, such as conferences, journals, or using new mediums such as microblogging. While electronic papers have played and continue to play a primordial role in the dissemination of research results, researchers now recognize that they are by no means sufficient to communicate and share research results. Indeed, the hypothesis investigated during the research, the experiment designed to assess the validity of the hypothesis, the process (workflow) used to ran the experiment, the datasets used and the results produced by the experiment, and the conclusions drawn by the scientist, are all elements that may be needed to understand, assess the claim, or be able to re-use the results of previous research investigations. Note also that, increasingly, scientists use workflows as a means to design and automate their data-intensive experiments. A workflow can be defined as a directed graph in which the nodes are data transformations that are implemented by software programs, e.g., web services, and the edges specify data flow dependencies. For example, we observe that researchers from modern sciences, notably from the life sciences, have recently started to publish and share, in addition to the paper describing their investigation, the workflow implementing their experiment using public web portals such as myExperiment[6].

As a step in this direction, we present research objects as an abstraction for communicating, sharing and reusing research results. As well as the paper describing the contribution of the scientists, a research object bundles additional information about the hypothesis the scientist investigated, the workflow implementing their experiment, the dataset they used, and the results obtained by running the workflow. The research object also contains annotations that describe all these elements with the view to facilitate the discovery and understanding, and therefore, the reuse of existing research objects in the context of new research investigations.

There are many open source software products that are publicly available for managing electronic papers, e.g., the Open Journal Systems[1], the e-Publishing Services[2]. While such tools provide useful capabilities to track electronic papers from their submission, through to their revision, to their publication, given the rich nature of research objects, compared electronic papers, we believe that the management of research objects poses additional challenges that need to be addressed. In this paper, we identify a set of research problems that together aims together to enable the management of research

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.

[1] http://pkp.sfu.ca/?q=ojs
[2] http://www.atypon-link.com

objects. We also point out the important role that end-users and automation techniques can play to enable the scalable management of research objects.

Accordingly, the paper is structured as follows. We begin by presenting the nature and spectrum of possible research objects. We go on to present the operations necessary for ensuring the management of research objects. We then underline the role that end users (i.e., scientists) involvement and automation techniques can play in scaling the management of research objects, before we close the paper.

RESEARCH OBJECT SPECTRUM

As mentioned earlier, a research object is expected to bundle different kinds of artifacts: paper, hypothesis, workflow implementing the experiment that assess the hypothesis, data sets used and/or produced, conclusions derived by the scientists. That said, it will be unrealistic to expect that all scientists will provide all these elements when publishing their research objects. Also, depending on the investigation communicated through the research object, it may or may not be useful to include certain kinds of elements. For example, if the research object is used to disseminate a vision or a survey on the state of the art, then it may not make sense to include experiments or data sets.

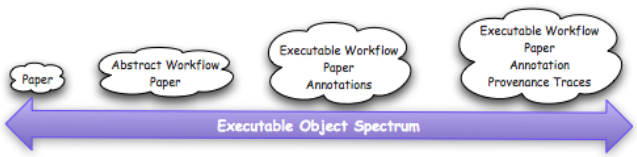

Figure 1. research object Spectrum

To illustrate the above aspect, Figure 1 depicts a spectrum illustrating the different kinds of research objects. At one end of the spectrum, left hand side, the research object is represented by a paper. As we progress to the other end of spectrum, the research object is enriched to include elements such as the workflow implementing the experiment, annotations describing, e.g., the experiment implemented and the hypothesis investigated, and provenance traces of past executions of the workflow [5].

RESEARCH OBJECT MANAGEMENT

Research objects, be they simple electronic papers or rich objects that contain information about the experiment and workflow used, research objects need to be managed to allow for their creation, curation, and sharing. We outline in this section, the functionalities that need to be provided for these purposes.

Preservation

Just like with electronic objects, research objects need to be preserved over time, to ensure their availability and the accessibility of the elements they are composed of. The preservation of a research object, however, poses additional challenges compared with the preservation of electronic papers. In particular, users may not to be able to execute the workflow [2] implementing the experiment described within a research object. This may be due, e.g., to the unavailability of the software components used within the workflow. Therefore, there is a need for a means to ensure that such workflows are preserved and can be executed years after the research object publication, which can be challenging [1].

Annotation

In order to allow scientists to understand and, ultimately, reuse a research object, annotations describing the elements of the research object should be provided. In particular, users should be able to understand how the elements of a research object fit together, understand the steps that constitute the workflow within the research object, and how the results produced by such a workflow enabled the derivation of the conclusions drawn by the scientist who published the research object.

Versioning

A user (or author) may wish to add new elements or modify existing ones within a research object. Such operations may yield the creation of a new version of the research object in question. It follows then that there is a need for a mechanism for versioning research objects to support the creation of a research object, maintain information about the different versions of a research object as well as ensuring their integrity. There is a plethora of version management systems, e.g., Subversion[3], Git[4] and Mercurial[5]. Such system may, however, need to be adapted to the need for versioning research objects, e.g., to ensure the links between input and output data on one hand and workflows on the other hand are valid.

Editing research objects

Users need to be able to edit research objects by aggregating documents, data and workflows together. Such users are not necessarily information technology experts. To support them in editing research objects, there is a need for tools, e.g., a workbench, that allows them to fetch and aggregate existing data, to design methods, e.g. workflows, to enact those methods, to store the results obtained as well as any metadata that the user may wish to add with the purpose of facilitating research objects discovery and reuse.

Provenance Management

Provenance plays a key role in understanding the dependencies between the elements that constitute a research object and the dependencies between the elements of different research objects. For instance, to assess the outcome claimed within a research object, evaluators may need to trace back the data that contributed to that outcome, e.g., the evaluator may want to know the data inputs used to produce a given workflow result. Provenance is a key ingredient to other activities, e.g., to understand, compare and debug research objects. Therefore, there is a need for collecting provenance of the elements that compose research objects and the traces of

[3]http://subversion.tigris.org
[4]http://git-scm.com
[5]http://mercurial.selenic.com

workflow executions. As well as logging provenance information, support for browsing and querying provenance [4] is required to facilitate the tasks users have at hand.

Browsing and Querying research objects
Users should be able to browse research objects using imprecise queries, e.g., keyword queries, as well as precise queries that specify the properties of the research objects to be retrieved, e.g., predicated queries. For example, a user who is interested in gaining knowledge of specific domain, say Astronomy, will be interested in browsing research objects that tackles that domain. In doing so, the user may want to examine the components of a research object exploiting intra-references that aggregate those components. The user may also explore other research objects by exploiting associations that connect research objects, e.g., to consult previous versions of a research object or to examine the research objects that make use of the research object s/he is examining. On the other hand, a power user, may be interested in locating research objects with specific properties. An example of a query such user may issue is *give me the research objects with a workflow that consume the same data inputs as a given workflow*.

USERS + INCREMENTALITY + AUTOMATION = SCALABILITY

Many of the functionalities presented in the previous section pose scalability issues. In particular, the annotation of research objects can get tedious as the number of research objects published or shared grows. In addition, retrieving research objects may be tricky when the number of research objects that needs to be examined to evaluate user's query is large. We believe that user involvement in the process of curation and annotations of research objects, together with tools for the automatic creation and indexing of research objects can be useful in facing and attenuating scalability issues.

Curating and Annotating Research Objects Using the Crowd
Assigning the operation of creation, publication and annotation of research objects to few information technology experts will quickly yield scalability issues. Instead, we believe that enlisting the crowd of scientists and delegating to them the responsibilities of creation, curation and annotation of research objects will allow to effectively deal with the large number of research objects.

Incrementality
A research object does not need to be fully annotated before its publication. Instead, we anticipate that the annotation process will be incremental: as research objects get (re-)used, new annotations will be added based on the experience of the users, and other annotations will be updated to reflect the current state of the research object. For example, if a dataset or a software component that is used in the experiment reported on in the research object are no longer available, then the user can update the annotation describing the state of the data set or software component in question.

Automation
Users may have to query a large population of research objects. For example, to identify the research objects that are similar to a given one, the user may need to query all known research objects. Accessing and querying a large population of research objects is likely to give rise to performance issues. Indexing support [3] is a mechanism that can be used to overcome this issue. Research objects are rich structures that bundle elements of different types and reference other research objects. Therefore, the indexing support used to facilitate access to such structures should cater for the richness (and therefore, the heterogeneity) of research objects in terms of contents.

CONCLUSIONS

As research practice evolves, we expect the emergence of new form of publications that are richer than the current electronic paper, which we term research objects. We have identified in this position paper the issues that need to be addressed to manage research objects, and outline the primordial role that end users involvement together with automation techniques, such as indexing, can play in dealing with scalability issues. We are investigating the issues identified, and the solutions outlined in the context of the Wf4Ever[6], a digital library European project.

ACKNOWLEDGMENTS

Wf4Ever is funded under the Seventh Framework Programme of the European Commission, project reference 270192 in the Digital Libraries and Digital Preservation area (ICT-2009.4.1). We are grateful to all our collaborators in theWf4Ever project.

REFERENCES

1. Belhajjame, K., Goble, C. A., Soiland-Reyes, S., and Roure, D. D. Fostering scientific workflow preservation through discovery of substitute services. In *IEEE eScience conference*, IEEE CS (2011).

2. Deelman, E., Gannon, D., Shields, M., and Taylor, I. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems 25*, 5 (2009), 528–540.

3. Dong, X., and Halevy, A. Y. Indexing dataspaces. In *SIGMOD Conference*, C. Y. Chan, B. C. Ooi, and A. Zhou, Eds., ACM (2007), 43–54.

4. Missier, P., Paton, N. W., and Belhajjame, K. Fine-grained and efficient lineage querying of collection-based workflow provenance. In *EDBT*, I. Manolescu, S. Spaccapietra, J. Teubner, M. Kitsuregawa, A. Léger, F. Naumann, A. Ailamaki, and F. Özcan, Eds., vol. 426 of *ACM International Conference Proceeding Series*, ACM (2010), 299–310.

5. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P. T., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. G., and den Bussche, J. V. The open provenance model core specification (v1.1). *Future Generation Comp. Syst. 27*, 6 (2011), 743–756.

[6] http://www.wf4ever-project.org

6. Roure, D. D., Goble, C. A., and Stevens, R. The design and realisation of the $\text{my}_{\text{experiment}}$ virtual research environment for social sharing of workflows. *Future Generation Comp. Syst. 25*, 5 (2009), 561–567.