SOLE: Linking Research Papers with Science Objects

Quan Pham[1], Tanu Malik[2], Ian Foster[1,2], Roberto Di Lauro[3], and Raffaele Montella[3]

[1] Department of Computer Science, University of Chicago, Chicago, IL 60637
[2] Computation Institute, University of Chicago, Chicago, IL 60637
[3] Department of Applied Science, University of Napoli Parthenope, Napoli, 80143
quanpt@cs.uchicago.edu, tanum@ci.uchicago.edu

Abstract. We introduce Science Object Linking and Embedding (SOLE), a tool for linking research papers with associated *science objects*, such as source codes, datasets, annotations, workflows, packages, and virtual machine images. The objective of SOLE is to reduce the cost to an author of linking research papers with such science objects for the purpose of reproducible research. To this end, SOLE allows an author to use simple tags to delimit a science object to be associated with a research paper. It creates an adequate representation of the science object and manages a bibliography-like specification of science objects. Authors and readers can reference elements of this bibliography and associate them with phrases in the text of the research paper through a Web interface, in a similar manner to a traditional bibliography tool.

1 Introduction

Prior to the computational driven revolution in science, research papers provided the primary mechanism for sharing data. Papers summarized experiments involving small amount of data, derivations on that data, and associated methods and algorithms. Readers reproduced results through physical experimentation, hand calculation, and/or logical argument. But as scientific methods have become increasingly computational, involving large quantities of data, complex data manipulation and/or numerical simulation, and the use of large and often distributed software stacks, the paper often merely summarizes rather than describes the data and computation. A reader wanting to understand the paper fully requires access to further digital materials. Input and output data may be shared through websites and software may be made available through packages or virtual machine images. Such indirect linkages, however, are typically disconnected from the claims and the results in the paper–not allowing, for example, an equation in a paper to be mapped directly to its implementation.

With the growing emphasis on reproducible research, readers and reviewers increasingly often want to be able to assess the validity of findings and to verify results. Consequently, indirect linkages are not sufficient. Instead, we would like digital materials associated with the works described in a paper–what we term here the paper's *science objects*–to be closely associated with the text so that they can be accessed while reading the paper. Examples of such associations are linking a concept described in the paper to its implementation in source code; linking a description of a dataset to its metadata and digital object identifiers (DOIs); linking a figure in the paper to its derivation and

P. Groth and J. Frew (Eds.): IPAW 2012, LNCS 7525, pp. 203–208, 2012.

workflow, and linking data values referenced from another paper sources to the exact location in that other paper's PDF source.

While associating papers with science objects at this fine granularity may be desirable, the realization of this goal introduces at least three challenges. First, we face the need to transform each science object into a form amenable to linkage with a paper. In our work, this means that important classes and functions in source code files must be associated with URLs and that datasets must be recorded in registries that specify dataset locations and access methods. It also means that data analysis pipelines must be cast as workflows with appropriate wrappers and web services that specify inputs and functional forms, or alternatively associated with software on a adequately provisioned virtual image. A second challenge concerns the manner in which linkages are represented in papers. Using URLs to refer to science objects is often unwieldy, especially when an object is referenced multiple times. A third challenge relates to presentation: Clicking on a science object link should lead to adequate presentation to the user.

We demonstrate Science Object Linking and Embedding (SOLE), a system [15] that eases the process of linking research papers with science objects, such as source codes, datasets, workflows, and virtual images. Authors identify science objects with human-readable tags; SOLE converts each tagged science object into an associated linked data object with an associated URI. For ease of management, the tags, URIs, and accompanying representation are maintained in a registry: what is, in effect, a science object bibliography. To aid authors with the linking process, SOLE also provides a web interface that allows authors to associate groups of words in a research paper with one or more science object tags. Clicking a link in the text results in the display of an appropriate representation of the science object.

In the remainder of this article, we describe how SOLE works to ease author burden and demonstrate how SOLE has been used to enable reproducible research in the RDCEP project [4] in which research papers must be associated with several computational products.

2 Related Work

It is widely recognized that currently there is a lack of suitable incentives that attribute scientists for conducting reproducible research. However, the merits of conducting reproducible research are also widely accepted–it leads to scientific methods which have higher transparency and are more open. To improve transparency of research papers, some projects have demonstrated the concept of reproducible research paper by focussing on one or more aspects. Utopia [1] reproduces paper by associating concepts in paper with external annotations retrieved from an online meta data store. Annotations are publicly shared and readers can further comment upon them. Vistrails [9] creates reproducible papers by associating figures and results in the paper with executable components. It allows authors to publish workflows and associated provenance and hyperlink to it in a result or figure in the article. Sweave [14] and Dexy [6] are literate programming environments, which if adopted from the beginning of the scientific process can lead to papers with embedded source code and derived results. SOLE is particularly targeted towards authors for whom experimentation and writing research

papers continue to remain separate activities, but would like a less burdensome, yet efficient mechanism to associate their research papers *post-hoc* with the inputs and outputs of the scientific process. SOLE has similar goals as provided by websites such as RunMyCode.org [13], but provides the ability to associate a wider class of scientific inputs and outputs at a finer control.

3 SOLE

SOLE provides command-line tools for authors to create science objects, and a web-based interface for authors and readers to associate phrases in the paper with their corresponding science objects. To create a science object in SOLE, the author puts a tag on the science object with the following syntax:

$begin\ type\ name_1|\ldots|name_n$
[science object content]
end

in which $begin$ and end are delimiters of the tag, $type$ defines the kind of science object to create, and $name_1$ to $name_n$ are user-defined names. Thus, the same object can be tagged by more than one name. SOLE processes a tagged file and based on tag $type$ definitions creates a science object, which associates a set of metadata elements representing the object, including a reference to the object as a URI. Authors can place tags on source codes and text files to create SOs, such as source code snippets, annotations in PDFs, units of a workflow that can be executed on a given environment, and virtual machine images, described later in the section. After creation of a variety of science objects, authors/readers can load the paper in HTML format and associate phrases in the text with the name of the tag.

In SOLE, four kinds of science objects can currently be created and linked:

Language Objects. The author can import a local source code repository or a public domain code repository, such as Github, to create URIs for language objects defined in the source. Internally, SOLE uses Ctags [5] to create tags for language objects in a file, but appends a URI to the language objects. We have expanded the Ctags utility to allow users to tag more than one language object as a single object to be referenced in a paper. This is useful when an algorithm in a paper must be associated with multiple functions and data structure specifications defined in multiple files.

Annotated PDFs and Datasets. SOLE uses the Poppler library [12] to extract tagged annotations from PDF. The metadata of the tagged annotation includes the URI of the PDF, the exact location in the pdf where annotation was made, and the annotated text. Tagging of a dataset, should ideally retrieve the metadata associated with the dataset, associate a DOI, and provide some methods for data access. However, developing generic tags for all types of datasets is challenging since datasets exist in a variety of formats, and with a wide variety of access tools. Currently, an author can tag the metadata file of NetCDF and ASCII datasets to generate a corresponding URI on the entire dataset. We plan to make dataset tags more versatile by integrating them with DOIs.

Web Services. SOLE creates a web-service specification of functions specified by a user. The functions are delimited in the source file by inserting tags with the workflow

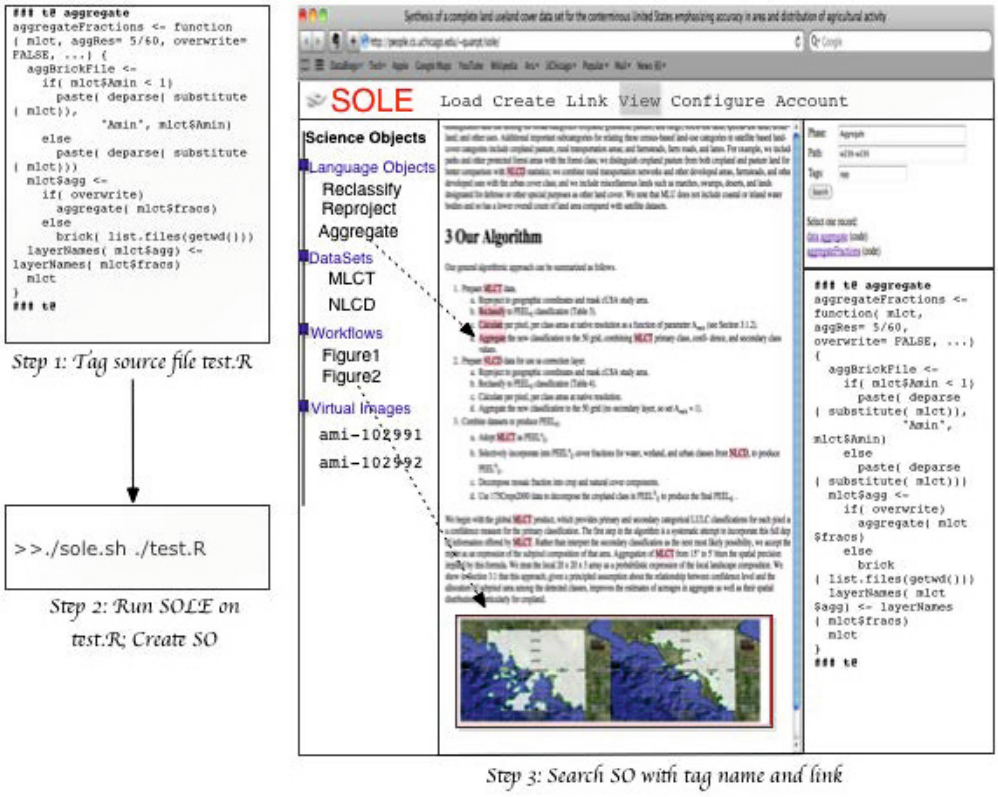

Fig. 1. Figure 1 demonstrates the three easy steps that an author follows to create and associate science objects in SOLE. The author first tags the function *aggregateFractions* in their source code with the user-defined tag name "*aggregate*" (step 1); then runs SOLE as command line tool to create the necessary metadata (step 2); and finally associates the phrase "Aggregate' in the paper with the tag and views the specification of the object (step 3).

tag_type. SOLE creates workflow specifications as Galaxy tools, with description about inputs and outputs. Galaxy provides an open, web-based platform for specifying tools and running computational experiments as workflows[10]. Each function is automatically wrapped as an appropriate Galaxy tool definition and hosted on the web-server instance connected with Galaxy. Authors can further specify if web services should accept user specified parameters and types of data.

Virtual images. Authors can also create packages of a source directory, using different package managers, and then with a single click deploy those packages on a virtual machine hosted on a cloud, and obtain a URI that includes machine ID and parameters for the package to be executed. To conduct this operation, SOLE must be configured with the user's account on a cloud infrastructure such as Amazon. SOLE uses a configuration file to specify the package and deploys on the image using recipes in provisioning tools.

The science object URI and its tag is stored in FluidInfo [7], a key-value data store that stores tags for a variety of data objects and provides a simple query language to allow users to search the datastore for specific tags and tag-values.

4 Demonstration Scenario

The Center for Robust Decision making on Climate and Energy Policy (RDCEP) [4] is a collaborative, multi-institutional project that aims to improve the computational models needed to evaluate climate and energy policies and to make robust decisions based on outcomes. Sharing science objects in the form of data, tools, and software is critical; it enables scientists to compare models and to build more accurate models. Currently in RDCEP science objects are shared through a web site. Our demonstration scenario consists of two documents [2,11] produced within the Center which we link with their respective science objects, using SOLE

Scenario 1: To reproduce the first document [2] (a master's thesis; see also the associated paper [3]), the author must associate the text and embedded figures with science objects that include datasets, algorithmic descriptions, computational analysis workflows, and workflow executions. The author tags each science object to create web accessible resources in the form of HTML fragments and web services. The resulting object representations are maintained in the SOLE database.

Scenario 2: To reproduce the second paper [11], the author must associate descriptions in the paper with a set of data values, each of which is embedded in another research paper. We demonstrate that authors can insert an annotations on the PDF, tag it, then use SOLE on PDF files to generate URIs on tagged annotations, and and finally associate with phrases in the research paper.

5 Conclusion

SOLE eases the management and creation of digital objects associated with scientific experiments and associating the objects with research papers. Demonstrated in the domain of policy science, SOLE uses general features and interfaces such as tagging and Galaxy. With minimal effort it can also be applied to other domains such as biology, astronomy, and the geosciences. SOLE is currently under development and will be released to a broader audience at a later date. In the future, we plan to interface SOLE to Globus Online [8] to enable authors to create a richer reproducible environment.

Acknowledgements. We thank Neil Best for the insightful discussions on conducting reproducible research. This work was supported in part by the Center for Robust Decision making on Climate and Energy Policy, under NSF grant number 0951576.

References

1. Attwood, T.K., et al.: Utopia Documents: linking scholarly literature with research data. In: European Conference on Computational Biology, Ghent, Belgium (September 2010)
2. Best, N.: Synthesis of a complete land use/land cover data set for the conterminous United States emphasizing accuracy in area and distribution of agricultural activity, Master's Thesis, Northeastern Illinois University (2011)

3. Best, N., Elliott, J., Foster, I.: Synthesis of a complete land use/land cover data set for the conterminous United States. RDCEP, Working Paper 12-08, http://dx.doi.org/10.2139/ssrn.2051158
4. Center for Robust Decision making on Climate and Energy Policy (RDCEP), http://www.rdcep.org/
5. Ctags, http://ctags.sourceforge.net/ctags.html
6. Dexy, http://www.dexy.it/
7. FluidInfo, http://fluidinfo.com/
8. Foster, I.: Globus Online: Accelerating and democratizing science through cloud-based services. IEEE Internet Computing, 70–73 (May/June 2011)
9. Freire, J., Silva, C.T., Callahan, S.P., Santos, E., Scheidegger, C.E., Vo, H.T.: Managing Rapidly-Evolving Scientific Workflows. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 10–18. Springer, Heidelberg (2006)
10. Goecks, J., Nekrutenko, A., Taylor, J., The Galaxy Team: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11(8), R86 (2010)
11. Johnson, S., Moyer, E.: Feasibility of U.S. Renewable Portfolio Standards Under Cost Caps and Case Study for Illinois. RDCEP, Working Paper 12-07, http://dx.doi.org/10.2139/ssrn.1996621
12. Poppler, http://poppler.freedesktop.org/
13. Run My Code, http://www.runmycode.org
14. Leisch, F.: Sweave, Part I: Mixing R and LaTeX: A short introduction to the Sweave file format and corresponding R functions. R News 2(3), 28–31
15. Science Object Linking and Embedding (SOLE), http://www.ci.uchicago.edu/SOLE