An Approach to Enhancing Workflows Provenance by Leveraging Web 2.0 to Increase Information Sharing, Collaboration and Reuse

Aleksander Slominski

Department of Computer Science, Indiana University
Bloomington, IN, 47405, USA
{aslom}@cs.indiana.edu

Abstract. Web 2.0 promises a more enjoyable experience for creating content by users by providing easy-to-use information sharing and collaboration tools, and focusing on user-centered design. Provenance in Scientific Workflow Management is one kind of user-generated data that can benefit from using Web 2.0. We propose a simple set of Web 2.0 technologies that is simple to implement and can be immediately leveraged by scientific users. Using Atom Syndication Protocol to represent workflow state and its provenance users can easily disseminate their scientific results. Collaboration and authoring can be facilitated by using Atom Publishing Protocol and standard Web 2.0 blogging tools to publish and annotate provenance. Users can search stored provenance by using search engines. If search results are in standard Atom Syndication Protocol, for example when search engines support OpenSearch standard, then Atom feeds can be used to monitor provenance changes increasing the likelihood of discoveries. By using those Web 2.0 standards, the value of scientific provenance data increases by making it a natural part of growing a variety of user-generated scientific (and non-scientific) content.

Keywords: scientific workflow provenance, user-generated content, scientific notebook, atom syndication format, atom publishing protocol.

1 Introduction

Web 2.0 promises greater control over user-generated content and a more enjoyable experience for users by improving information sharing and collaboration capabilities, and focusing on user-centered design. Those benefits should not only be enjoyed by end-users but enterprise and scientific users as well.

There is an increasing interest in trying to leverage Web 2.0 benefits in science with the ultimate goal to create something that may be called "Science 2.0" [4]. Scientific Workflow Management (SWFMS) [14] is one of such areas of science that may benefit to leverage Web 2.0 (e.g. [6][7]). In particular, we believe that one of the areas where the biggest gains can be obtained from using Web 2.0 standards is to apply it to provenance in Scientific Workflow Management Systems. That will help to solve one of the biggest problems that scientists have when working on scientific workflows: how to collaborate and disseminate results. To collaborate scientists need

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, pp. 231–235, 2010.
© Springer-Verlag Berlin Heidelberg 2010

to have a shared environment (Web 2.0 is a good candidate) and to disseminate results they need to be able to not only send output files but allow other scientists to reproduce them (that is why provenance is important).

2 Web 2.0 and Scientific Workflow Provenance

Web 2.0 does not have one clear definition [16] but key characteristics of Web 2.0 around user-generated data can be identified [1]:

- Search: the ability to find useful information in ever- increasing amounts of data is a key feature of Web 2.0 for users. Making data and metadata generated in scientific workflows searchable by search engines is the simplest approach to accomplish it. Moreover, by using OpenSearch 1.1 [8] more customized search queries and results in formats required by scientists can be provided. For example, search results in standard Atom Syndication Protocol [10] can represent provenance and its metadata (as in examples below).
- (Hyper) Links are the key ingredient to the Internet experience. If scientific content is not linked, it is as if did not "exist" on the Internet. The ability to cite and reference is an integral part of scientific process and scientific papers – the same concept in Science 2.0 is implemented by using links.
- Authoring has been the enabling factor in making Web 2.0 successful. In particular, blogs provide easy-to-use platforms that ordinary users can leverage with minimal computer-science experience. Scientific Workflows are managed by scientific users (such as scientists) that, in majority, prefer to concentrate on their science than to become proficient in computer-science. Leveraging well-tested Web 2.0 authoring tools in scientific environments, therefore, can lead to quicker dissemination of results as well as easier sharing and collaboration.
- Tags are a very easy-to-use tool for organizing information without the need for users to learn and understand taxonomies. Tags provide bottom-up taxonomy and should significantly help scientific users organize quickly increasing amounts of data produced in science.
- Syndication and publication of content is made easy with standards such as The Atom Publishing Protocol [9]. The pull model of publishing works well to broadcast scientific results to interested subscribers and facilitate collaboration.

We believe that the Scientific Workflow Provenance can benefit from network effect by making it a natural part of a variety of Web 2.0 user-generated content. To achieve this there are several proposed solutions. A leading approach is to build on a larger framework of Semantic Web and in particular Linked Data [11] with HTTP URIs and combination of RDF and SPARQL as machine readable data format and query language. However Semantic Web requires a layer of sophisticated middleware to achieve its goals - it is much more than a simple extension of World Wide Web and using it for provenance require additional work [13][15].

As a simple alternative to full stack of Semantic Web one could use The Atom Syndication Format [10] as it is an extensible entry data format that can embed any XML data (including XHTML) and metadata about it (including links). That extensibility can be leveraged to encode scientific workflow provenance. In particular, the Open Provenance Model [2] may be good target to assure that provenance data is not

locked inside SWFMS. The query part can be fulfilled by search engines indexing pages generated from ATOM entries and feeds. For more targeted searches OpenSearch can be leveraged.

Even though ATOM provides only a subset of capabilities available in fully featured Semantic Web solution, ATOM entries can be easily transformed into RDF and can be fully integrated into evolving Semantic Web middleware.

3 Using the Open Provenance Model with ATOM

In the Open Provenance Model (OPM) [2] there are several types of nodes, including Artifact, Process, Agent and OPM nodes and edges:

- Artifact is an "immutable piece of state, which may have a physical embodiment in a physical object or a digital representation in a computer system";
- Process is defined as an "action or series of actions performed on or caused by artifacts, and resulting in new artifacts";
- Agent is a "contextual entity acting as a catalyst of a process, enabling, facilitating, controlling, or affecting its execution";
- OPM also describes relations between nodes, edges in a graph. Typical edges are "used," "wasGeneratedBy," and "wasDerivedFrom."

OPM nodes and edges could be naturally represented as ATOM entries, edges translated to links in ATOM entries, and an OPM graph becomes an ATOM feed. There are already existing proposals to do it. In this paper we show a very simple encoding of OPM into ATOM. Each ATOM entry has atom:link to give URL to retrieve HTML representation and a unique atom:id that works well as an graph node ID. Moreover, any part of XML representation of OPM (such as [3]) can be embedded:

```
      <entry>
        <title>Workflow 1 in Foo Version  1.1 </title>
        <link href="http://example.org/foo/1.1/w1 "/>
        <id>urn:uuid:...-888888888</id>
        <updated>2010-03-13T17:00:03Z</updated>
        <summary>Workflow 1 was executed by system Foo
Version  1.1.</summary>
        <opm:Agent id="urn:uuid:...-888888888" />
      </entry>
      <entry>
        <title>Input file bar.dat</title>
        <link href="http://example.org/f/bar.dat"/>
        <id>urn:uuid:...-9999999999</id>
        <updated>2010-03-13T17:01:03Z</updated>
        <summary>Input file bar.dat</summary>
        <opm:Artifact id="urn:uuid:...-9999999999" />
        <category scheme="http://..." term="bar" />
      </entry>
```

User actions will naturally correspond to blog postings, with the "agent" doing work (or becoming an author)

```
<entry>
  <title>Running Xyz Processing</title>
  <link href="http://example.org/w/AAAA "/>
  <id>urn:uuid:...-AAAAAAAA</id>
  <updated>2010-03-13T17:11:03Z</updated>
  <summary>Xyz processed bar.dat </summary>
  <opm:Process id="urn:uuid:...- 888888888" />
  <oprel:used id="urn:uuid:...-9999999999" />
  <link rel="http://.../opm/rel#used"
     href="http://example.org/f/bar.dat" />
</entry>
```

4 Use Case Scenario

To illustrate how scientific workflow users can benefit from Web 2.0 integration, we describe a scenario where Alice is monitoring workflows that are run by Bob. Bob either manually or by using a job scheduler runs large weather forecast workflows using his custom code. The workflow system Foo used by Bob is running a process that Bob designed and that leverages Bob's experience. In particular, he has designed a special Xyz processing used in his workflows. For each new process started, the Foo system creates a new ATOM feed (OPM graph) and publishes workflow progress as ATOM entries to this feed. The system also publishes ATOM entry to public ATOM feed when a new process starts. Alice is using specialized software to monitor this feed (it could be a slightly modified commercial blog reader). If Bob used PubSub-Hubbub protocol [12] then Alice could get near-instant notifications about changes in Bob workflows.

When results of Xyz processing are published, they are automatically downloaded to Alice's desktop (standard function of blog software) and analytics code is executed (additional software required). The analytics could also be a workflow process that publishes results as ATOM feed so Bob (and other scientists) can monitor it and verify provenance of results. When analytics detects interesting conditions (such as a strong possibility of a tornado), an alarm is published to high priority ATOM feed to which Alice is subscribed (by email, SMS, etc.).

When Alice finds something interesting about Bob's results, she can publish it to her blog or post a comment to Bob's workflow feed to let Bob know that his workflow produced something that may be incorporated in Alice's future publications.

5 Summary

We hope that we demonstrated that provenance publishing and collaborating on scientific results can be facilitated by using Atom Publishing Protocol and standard Web 2.0 blogging tools. With easy migration path for data stored in ATOM format to future Semantic Web and Provenance standards additional benefits can be leveraged in the future as Web 2.0-related technologies mature and become attractive to scientists.

References

1. McAfee, A.P.: Enterprise 2.0: The Dawn of Emergent Collaboration. MITSloan Management Review 47, 21–28 (2006)
2. The Open Provenance Model Core Specification (v1.1), http://eprints.ecs.soton.ac.uk/18332/1/opm.pdf
3. http://github.com/lucmoreau/OpenProvenanceModel/blob/master/opm/src/main/resources/opm.1_1.xsd
4. Waldrop M.: Science 2.0: Great New Tool, or Great Risk? In Scientific American (May 2008)
5. Harrison, A., Taylor, I.: Web enabling desktop workflow applications. In: SC-WORKS 2009 (2009)
6. De Roure, D., Goble, C.: myExperiment: A Web 2.0 Virtual Research Environment for Research using Computation and Services. In: Workshop On Integrating Digital Library Content with Computational Tools and Services at JCDL 2009, Austin, Texas, USA (19-06-2009)
7. De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D., Newman, D.: myExperiment: Defining the Social Virtual Research Environment. In: 4th IEEE International Conference on e-Science, Indianapolis, Indiana, USA, December 7-12 (2008)
8. OpenSearch 1.1 Specification, http://www.opensearch.org/Specifications/OpenSearch/1.1
9. The Atom Publishing Protocol. Internet Official Protocol Standards, RFC 5023 (October 2007), http://tools.ietf.org/html/rfc5023
10. The Atom Syndication Format. Internet Official Protocol Standards, RFC 4287 (December 2005), http://tools.ietf.org/html/rfc4287
11. Berners-Lee, T.: Linked data, http://www.w3.org/DesignIssues/LinkedData.html
12. PubSubHubbub Core 0.3 – Working Draft, http://pubsubhubbub.googlecode.com/svn/trunk/pubsubhubbub-core-0.3.html
13. Hartig, O.: Provenance Information in the Web of Data. In: Proceedings of the Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW), Madrid, Spain (April 2009)
14. Yu, J., Buyya, R.: A Taxonomy of Workflow Management Systems for Grid Computing. Technical Report GRIDS-TR-2005-1, Grid Computing and Distributed Systems Laboratory, University of Melbourne (2005), http://www.gridbus.org/reports/GridWorkflowTaxonomy.pdf
15. Moreau, L.: The Foundations for Provenance on the Web. Foundations and Trends in Web Science, http://eprints.ecs.soton.ac.uk/18176/ (submitted)
16. Sharma, P.: Core Characteristics of Web 2.0 Services. (Published 28 November 2008), http://www.techpluto.com/web-20-services/