



RESOURCESYNC

ResourceSync Press Release - December 14, 2011 - New Standard Will Improve Availability and Timeliness of Repository Content

The National Information Standards Organization (NISO) and the Open Archives Initiative have been awarded a \$222,000 grant for a joint project to develop a new open standard on the real-time synchronization of Web resources. Increasingly, large-scale digital collections are available from multiple hosting locations, are cached at multiple servers, and leveraged by several services. This proliferation of replicated copies of works or data on the Internet has created an increasingly challenging problem of keeping the repositories' holdings and the services that leverage them up-to-date and accurate. As we move from a Web of documents to a Web of data, synchronization becomes even more important: decisions made based on unsynchronized or incoherent scientific or economic data can have serious deleterious impact.

"This proposal is an outgrowth of the issues exposed in the context of the Memento project that developed a protocol for uniformly accessing time-stamped resource versions on the Web," explains Michael L. Nelson, Associate Professor, Old Dominion University Computer, and a principal investigator on the Memento project. "We have assembled a stellar core team to devise the standard from the Sloan grant. It includes people that have worked on a variety of information interoperability efforts such as Memento; the OAI Object Reuse and Exchange (ORE), a protocol for describing aggregations of Web resources; Open Annotation, a resource-centric annotation framework; and the DSNotify change detection framework for Linked Data."

"The OAI Protocol for Metadata Harvesting (PMH) 2.0 specification can be used to effectively synchronize the metadata about the resources," states Carl Lagoze, Associate Professor, Cornell Information Science, and OAI Executive, "but synchronizing the resources themselves was never specified. Although some resource synchronization methods exist, they are generally ad hoc, arranged by the individuals involved, and cannot be universally deployed."

"As we started to explore this problem domain, it became increasingly clear that an interoperable, efficient and lightweight mechanism to support synchronizing resources at scale is missing from the Web infrastructure;" states Herbert Van de Sompel, Scientist, Los Alamos National Laboratory, OAI Executive and a principal investigator on the Memento project. "Candidate technologies that can help to tackle the problem exist, but they need to be put together and profiled in an appropriate manner."

"We expect this new standard will save a tremendous amount of time, effort, and resources by repository managers through the automation of the replication and updating process," states Todd Carpenter, NISO Managing Director. "The end result will be to increase the general availability of content in Web repositories and alleviate the variety of problems created by outdated, inaccurate, superseded content that exists on the Internet today."

A new work item for the project has been approved by the NISO Voting Members; see http://www.niso.org/apps/group_public/document.php?document_id=7512. Stay in touch by subscribing to resourcesync@list.niso.org.

An Initial ResourceSync Experiment

In order to explore the ResourceSync problem, a straw man resource synchronization approach was formulated and tested. The approach consists of:

- A Source pushing change notifications to Destinations using XMPP PubSub as a carrier protocol – Push CN in Figure 1.
- A tweet-like change notification language, deliberately chosen to be writeable and readable by both machine and human agents – Figure 2.
- A Destination pulling the entire resource about which it received a change notification from the Source – Pull CT in Figure 1.

The approach was tested to synchronize two Destinations with a Source that is a LiveDBpedia instance operated at the Los Alamos National Laboratory.

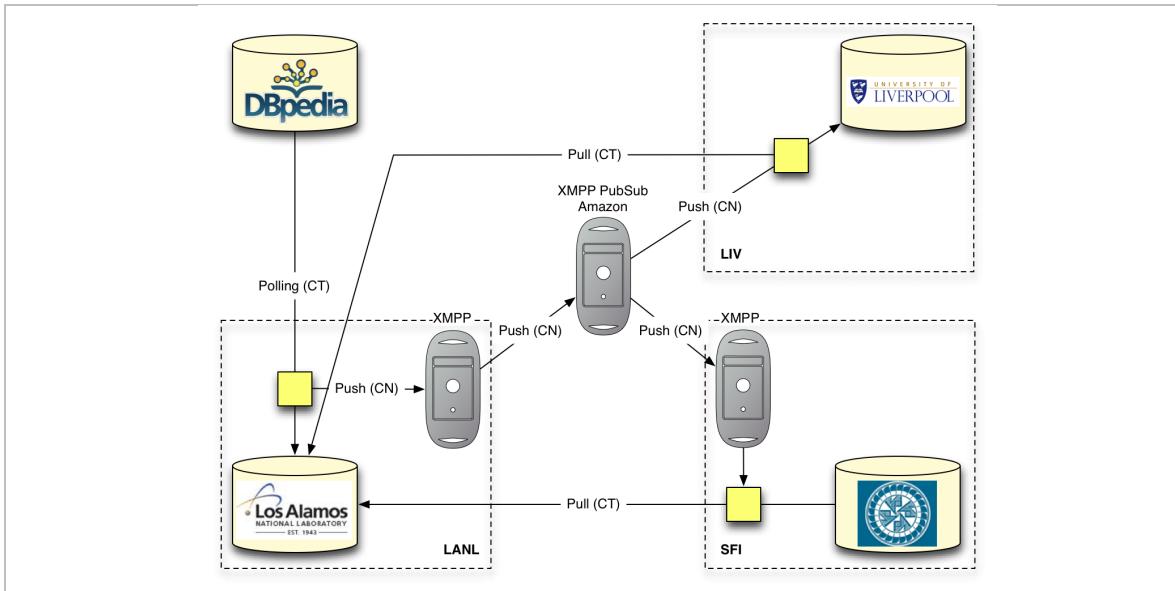


Figure 1: Two Destinations (right) synchronize with a LiveDBpedia instance (bottom left) by acting upon change notifications to which they subscribe.

```
http://megalodon.lanl.gov/dbpedia/data/Paris updated at="2012-03-05T19:54:39Z"
#dbpedia $resync
```

Figure 2: A change notification aka a bleep indicates that a resource was updated. Upon receipt of the bleep, the Destination can retrieve the updated resource.

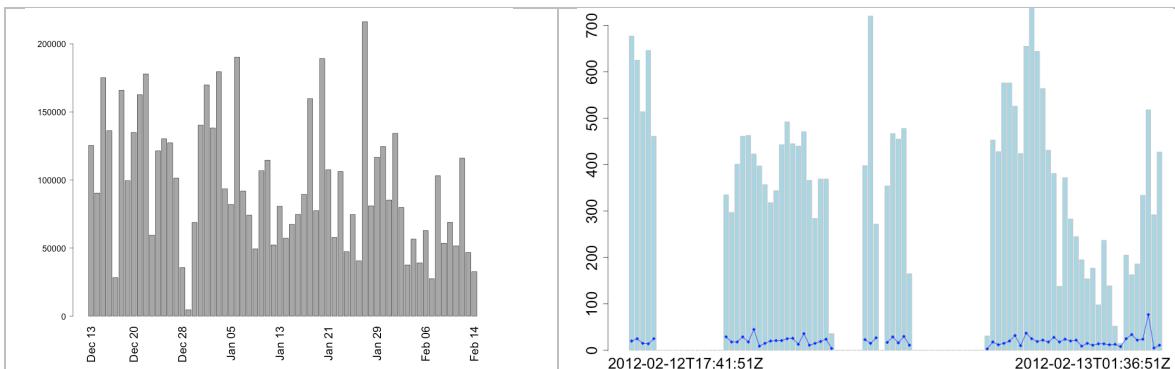


Figure 3: Left: An average of 2 change notifications per second were sent. Right: Destinations easily synchronized changed resources under this notification load. The bars are bleeps received in a time slot, whereas the line indicates the max amount of resources in the processing queue in the time slot; queues were always empty at the end of the time slot.