Motivated by the need to annotate data stored in authoritative data stores, but displayed to users and services capable of improving the data quality in many places distributed and remote from the authoritative stores, we have designed and implemented a prototype Filtered Push (FP) network in the domain of botanical collections built over the open source Map-Reduce platform Hadoop and the Java Messaging Service ActiveMQ including network clients in the open source biological collections management system Specify 6 and the GBIF Integrated Publishing Toolkit. The FP software is a domain-neutral framework for originating, distributing, and analyzing record-level annotations.

Find Duplicates

The problem: I'm entering data from a herbarium sheet, and I'd like to get the data someone else has already captured on a duplicate of this sheet in another institution rather than retyping it all.

1) Query: FindDuplicatesMessage

2) Triage plans and executes a series of actions in response to the Query:
   I) Perform fuzzy match on Collector=Demaree, Delzie and CollectorNumber=38875 on local cache.
   II) Return results as potential duplicate sets.
   III) Send out to network to find further potential matches of similar material to load into local cache.

3) Rapid Response: ReturnMessage (containing multiple potential duplicates)

4) Data from selected duplicate is incorporated into local database

5) An annotation message is sent annotating the specimens as members of a duplicate set.

Make Annotation

The problem: My data appears somewhere on the web not connected to my database, if anyone makes an assertion about my data, I'd like to hear about it.

A) Capture the annotation in structured form in whatever remote place where the data are displayed.

B) Annotation is submitted as xml document to a SOAP service

D) Interested subscribers can pick up the messages from their queues.

C) Triage plans and executes an action on the message, sending to one or more message queues

To the left is an illustration of the basic FP architecture for transporting and distributing annotations, highlighting client interactions with the network.

The FP “Triage” module is the heart of an FP network’s dispatch of incoming annotations. It may invoke analysis tools (e.g. fuzzy matches on collector names to propose purtative duplicate specimen sheets) after which it may place the annotation and results of the analysis into a global annotation store, which we presently implement on top of the Apache Hadoop Map-Reduce framework. In addition to (or instead of) that, it may invoke the Annotation Service module which, among other things, is responsible for publishing notices about the annotation onto message queues using the Active MQ open source implementation of the Java Messaging Service standard interfaces. Finally, using JMS service calls, client side code can receive current messages on JMS message queues to which they subscribe. Clients may have subscribed (or recorded a desire to subscribe) to queues of interest to them (e.g. any queue of annotations about the Orchid family), or this subscription may have happened automatically if the client has launched a message whose type in our annotation Schema indicates it expects a reply.

Copyright © 2009 Zhimin Wang, Maureen Kelly, Robert A. Morris, Paul J. Morris. Permission is granted to copy, distribute and/or modify this document under the terms of the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.