

Filtered-Push: A Map-Reduce Platform for Collaborative Taxonomic Data Management

Zhimin Wang², Hui Dong², Maureen Kelly³, James A. Macklin³, Paul J. Morris¹,
Robert A. Morris²

¹ Harvard University Herbaria and Museum of Comparative Zoology

² Department of Computer Science, University of Massachusetts, Boston

³ Harvard University Herbaria

wangzm@cs.umb.edu, hdong1@cs.umb.edu, mkelly@oeb.harvard.edu,
jmacklin@oeb.harvard.edu, mole@morris.net, ram@cs.umb.edu

Abstract

The Filtered-Push project aims to establish a cross-institutional infrastructure to help biologists (especially taxonomists) share and improve digitized natural history collection data via the exchange and management of specimen record annotations. Three challenges commonly confront the holders of data documenting specimens collected in the field: the identification of the organism and the annotation of records that arose from a single collection event but where parts of the organism have been distributed and are held as duplicates in multiple institutions; the quality control of new annotations [2]; and, more generally, the dissemination of annotations of specimen records, whether or not representing duplicate specimens. Addressing these can accelerate the rate of digital capture of data from paper records (such as handwritten labels attached to pinned insects or pasted on herbarium sheets with dried plants) and provide mechanisms for the global community of biologists to improve the quality of the data.*

1. Introduction

The biological collections community has put a great deal of effort into digitizing existing specimen records in natural history museums and other biological collections worldwide since the 1970s. The specimens in these collections are the basis for our knowledge of what species of organisms exist in nature and where they occur. Many millions of collection records have been digitized and are available in networks such as those supported by the Global

Biodiversity Information Facility (Edwards et al [12]; Bisby et al. [13]) and the Ocean Biogeographic Information System (Grassle and Stocks [14]). Many more specimen records, however, remain to be captured, particularly in the disciplines of entomology and botany, where historical practices haven't included ledgers, card files, or other paper records in forms that allow efficient and rapid data capture without examining every specimen in the collection (Beccaloni et al [15]). In the NSF-sponsored Herbarium Networks Workshop held in 2004 [17], 25 participants agreed on the goal to make all botanical specimen information in U.S. collections available online by 2020. However, the estimated number of specimen in U.S. herbaria is about 95,000,000 (Beaman, R. pers. comm.), of which only 5% have been captured over the last 30 years. Roughly it would take 9500 person years to complete the job, assuming it takes 10 minutes to finish the data input including geo-referencing for every specimen record.

One way to tackle this problem is to automate the capture process through image and natural language processing, which is underway with NSF support to R. Beaman and his colleagues (<http://www.herbis.org>). Collaboration and pooling of effort, particularly for geo-referencing collecting locality data, has been very successful in the vertebrate zoology collections, e.g. Stein and Wiczorek [16]. The practices of botanists allow us to approach the problem of costs and scale of data capture from this perspective of collaboration. Botanists, largely uniquely to their discipline, normally distribute specimen duplicates over institutions. Pieces of the same individual plant and associated data are distributed to more than one herbarium. Rabeler and Macklin [1] found that a large (perhaps 50%) portion of the 90,000,000 specimens remaining to be captured are duplicated at least once. This finding leads

* This research was partially funded by NSF grant DBI-0646266

us to our solution: constructing a sharing network across institutions, which will dramatically reduce the data capture costs by effectively avoiding repeated effort.

In fact our design also considers two other important factors for data capture and sharing: by incorporating more information and authority files over the network, the software can more easily detect possible data errors thus reducing the time a human requires to do quality control; We also propose a filtered-push approach to reconcile the global and local views of related data, which give users a configurable interface to respond to annotation changes that could affect local data.

2. Relevant works

Many current electronic mechanisms for dealing with specimen record annotations typically involve unscalable technologies such as e-mail followed by manual insertion of annotation into the recipient's database. For example in 1992 Australian herbaria began distributing data for exchanged specimens using the HISPID3 standard fields over email [3]. HISPID provides for wholesale importation of specimen records when duplicate specimens are distributed, a common practice among botanists. This is designed to reduce the cost of specimen record management at the time of distribution of duplicate specimens [4]. However, it is not designed to discover, examine, and accept, reject, or comment upon existing annotations of duplicate specimens. To our knowledge, no current approaches support mechanisms for discovery of existing annotations relevant to a given specimen record, nor any form of interested-party notification beyond email and RSS, nor are any capable of error detection beyond simple syntax checking. For example, the approach we will describe can analyze whether a collector name or a geo-location associated with a specimen record is likely to be a misspelling and if the specimen is part of a duplicate set.

3. System Design

Our approach to the problems proceeds in two steps: the first is to create a uniform global view of data among participating data nodes that includes definition of a common vocabulary and communication messages; the second is to construct a pluggable execution framework, which will carry and route messages, provide support for the network and message receiving agents to act upon and respond to messages based on local or global semantics.

3.1. Common Vocabulary and Supported Message Formats

The current underlying data model is a duplicates-oriented global view of specimen data of participants. Through this view, local changes can be provided globally and new global annotations can be applied to local copies. We currently use a global vocabulary based on the ABCD specimen record standard [5]. However, in the future we may adopt RDF (Resource Description Framework, <http://www.w3.org/RDF/>) mapping to support more extensible and semantically richer features.

As to messages, we have an XML schema type `MessageType`, which contains a header (common to all messages) and a body (which can be extended to support specific message types). In the message header, the main elements are information about senders, recipients and message types. Message passing adheres to a publication/subscription model in which nodes or individual users can subscribe to classes of messages based on their type, or on issues of special interest in this problem, such as the taxonomic group entailed in the message, particular institutional collections impacted by it, geographic locations at which the subject specimens were collected, etc. Currently messages that we support through extension of the message body can be divided into two categories: *pull* messages, i.e. those for which a response is expected, and *push* messages, which do not. The former include queries seeking existing specimen annotations and, those seeking duplicates whose existence may not yet be known to the network. Push messages include new annotation proposals such as new information, which for example, may include a new taxonomic identification about a specimen record, and/or correction of existing annotations.

3.2. Software Infrastructure Design

The infrastructure design of our framework comprises five components, dedicated respectively to: client API, network communication, network-to-local node adaptation (local-global vocabulary mapping), network node adapters which map the logical overlay network to the transport layer (usually http or email), and storage of network-wide information, e.g. subscription data. All these modules are glued together through well defined interfaces, which also give us the flexibility to change the underlying implementation of each module. Figure 3.1 illustrates the connection between different components.

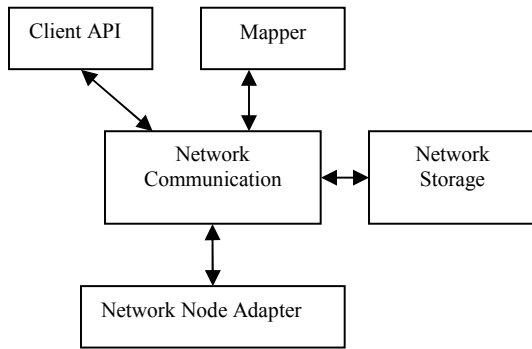


Figure 3.1

Messages inside the network communication layer and data stored in network storage are represented with the common vocabulary. The mapping between local node data and network data is carried by the Mapper component. Network storage is used for global data such as those representing duplicates, and caches of data of local nodes. This global repository helps to retain knowledge that is independent of local nodes such as pending annotations.

4. System Implementation

Because of the distributed nature of the data nodes, we selected the Hadoop [6] map-reduce framework as a platform for the current implementation to take advantage of computation resources of the whole network and data locality. More particularly, even if it is a putative record of a duplicate specimen, a given record is wholly held in a single database, but deciding *if* it represents a duplicate requires launching a query into the network to find potential matches. Analysis computations such as the discovery of duplicates and identification of outliers for quality control, fit the map-reduce model well, in that one can understand the process as combining (reduce) results from local searches (map). The quality control application is of particular importance, since the overwhelming fraction of the estimated 1.5-3 billion natural history collection specimen records [8] [9] [10] are not digital born and many digitizing errors can be discovered by comparison of taxonomic or curatorial data between the holdings of different collections. Graham et al. [11] have observed that the increasing online availability of digitized specimen records opens many avenues for ecological modeling based on geolocated species occurrence, but carries special problems associated with data errors. In fact, the Apache open source

project Mahout [7] is dedicated to importing machine learning algorithms suitable to the map-reduce paradigm into Hadoop platform, which would bring a big potential for our system to be used as a data analysis platform for information lying in the specimen records.

Moreover, the Hadoop HBase distributed database provides high availability by its robust, transparent file replication architecture. Its column-oriented database structure is particularly suited to a global annotation store, from which local participants can accept or reject annotations based on local policies. In fact HBase provides not only the normal advantage of efficiently storing sparse data, but also some other nice features for annotations. For example, the HBase's columns are organized by column families, into which a new column can be inserted at anytime. This is exactly the dynamic structure we want for taxonomy data, where a new feature may be added to accommodate new kinds of annotation.

Error correction is only one of the things that give rise to changes in specimen data record annotation. As new species are discovered, as taxonomists' concepts of the scope of existing species changes, and as taxonomists recognize misidentified specimens in collections, a given specimen may be deemed to belong to a new or different species than originally assigned to it. Such a *taxonomic revision* or similar kinds of changes in the name of the place of collection (when not given as latitude and longitude data), not only give rise to further annotations of a record, but also motivate tracking the history of those annotations. We can use the mechanism that HBase provides by default for each column of a datum, which keeps its history by associating data of each column with timestamps.

Hadoop enables us to remain flexible about the exact design of network messages. At this writing the client view of messages is a publication/subscription ("pub/sub") model similar to that of the Java Message Service (JMS, <http://java.sun.com/products/jms/>). However, subscribers may wish to filter subscriptions to receive notice only of selected messages on the corresponding subscription queue, e.g. those corresponding to particular species, or species collected at a specific location, or by a particular collector, even if some subscription list has a broader purpose. Furthermore, as a message passes through the network, observer (software or human) agents may publish the message to subscription queues not given by the originator, thereby broadening the audience for that message beyond the intent of its originator.

Figure 4.1 is our current high-level UML component view of the system showing the components mentioned above. The JobManager,

MapReduce, and NetworkStorage are all implemented with Hadoop.

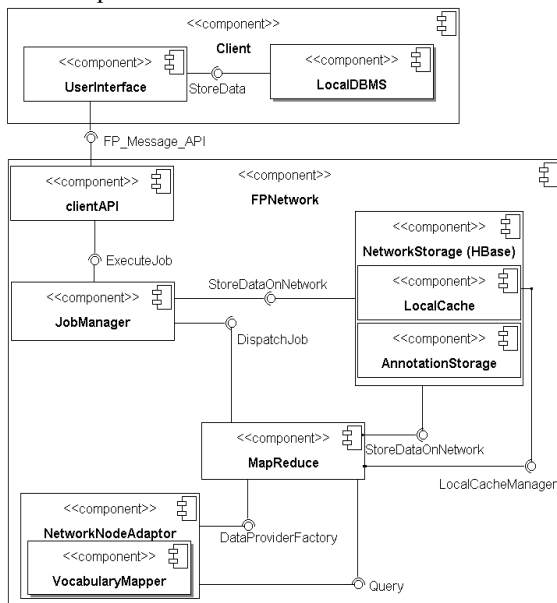


Figure 4.1

5. Conclusion

We have designed and implemented a system that imposes a strict separation between the message passing network and the computation models required to filter and respond to messages about data annotations in related databases. It will not only improve the efficiency of specimen data capture by finding duplicates of specimens and providing high confidence quality control, but also can provide a convenient platform for knowledge sharing among taxonomists and a cross-institutional data analysis engine for taxonomic and related research.

6. References

- [1] Macklin, J.A., R.K. Rabeler, and P.J. Morris. 2006. Developing a framework for exchange of botanical specimen data to reduce duplicative effort and improve quality using a 'filtered push'. <http://www.2006.botanyconference.org/engine/search/index.php?func=detail&aid=587> [Seen 2009 January 7]
- [2] Chapman, A. D. Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen, <http://www2.gbif.org/DataCleaning.pdf> p. 28. 2005.
- [3] Conn, B.J. Sharing of Accession-based botanical information – Reduction of Costs in Herbarium Data-entry in

Australia Using HISPID3 (Royal Botanic Gardens Sydney) [<http://www.rbgsyd.gov.au/HISCOM>] ;1998.

- [4] Neish P., Richardson B. and Whitbread G. New Standards from Old: reconciling HISPID with ABC, Biodiversity Information Standards (TDWG), Annual Meeting, Bratislava, 2007 http://www.tdwg.org/fileadmin/2007meeting/slides/Neish_HISPID_ABCD_abs235.ppt

- [5] TDWG, Access to Biological Collection Data - version 2.0.6; <http://www.tdwg.org/standards/115/>, 2005 (last visited 9/30/2008)

- [6] Apache Software Foundation. Hadoop <http://Hadoop.apache.org/core/2008> (last visited 9/30/2008)

- [7] Apache Software Foundation. Mahout <http://lucene.apache.org/mahout/>

- [8] Allison, A. "Biological surveys - new perspectives in the Pacific", *Organisms Diversity & Evolution*, Volume 3, Issue 2, 2003, Pages 103-110, ISSN 1439-6092, DOI: 10.1078/1439-6092-00065.

- [9] Howie, F. M. (1993): "Natural science collections: extent and scope of preservation problems." Pp. 97–110 in: Rose, C. L., Williams, S. L. & Gisbert, J. (eds.) *International Symposium and First World Congress on the Preservation and Conservation of Natural History Collections*, Madrid, Spain, 10–15 May 1992. Direction General de Bellas Artes y Archivos Ministerio de Cultura, Madrid.

- [10] Duckworth, W.D., Genoways, H. H. Rose, C. L. *Preserving natural science collections: chronicle of our environmental heritage*. Washington, DC: National Institute for the Conservation of Cultural Property, 1993.

- [11] Catherine H. Graham, Simon Ferrier, Falk Huettman, Craig Moritz, A. Townsend Peterson, *New developments in museum-based informatics and applications in biodiversity analysis*, *Trends in Ecology & Evolution*, Volume 19, Issue 9, September 2004, Pages 497-503, ISSN 0169-5347, DOI: 10.1016/j.tree.2004.07.006.

- [12] Edwards, J.A., M.A. Lane, E.S. Nielsen, 2000. Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science* 29(5488):2312-2314.

- [13] Bisby, F.A., J. Shimura, M. Ruggiero, J. Edwards and C. Haeuser. 2002. Taxonomy, at the click of a mouse. *Nature* 418:367

- [14] Grassle, J.F., and K. L. Stocks. 1999. A Global Ocean Biogeographic Information System (OBIS) for the Census of Marine Life. *Oceanography* 12(3):12-14

- [15] Beccaloni, G.W., M.J. Scoble, G.S. Robinson, A.C. Downton, and S.M. Lucas 2003. Computerizing Unit-Level Data in Natural History Card Archives. pp. 165-176 In M.J. Scoble ed. *ENHSIN: The European Natural History*

Specimen Information Network. The Natural History Museum, London.

[16] Stein, B. and J. Wiczorek. 2004. Mammals of the World: MaNIS as an example of data integration in a distributed network environment. *Biodiversity Informatics*, 1:14-22.

[17] R.K. Rabeler, Macklin, J.A. Herbarium Network in the United States: Towards Creating a Toolkit to Advance Specimen Data Capture. *Collection Forum* 2006; 21(1-2):223-231