Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 1614 – 1619

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Kurator: A Kepler Package for Data Curation Workflows

L. Dou[a], G. Cao[a], P.J. Morris[b,c], R.A. Morris[b], B. Ludäscher[a,1], J.A. Macklin[d], J. Hanken[c]

[a]*UC Davis Genome Center, University of California, 451 Health Sciences Drive, Davis, CA 95616, USA*
[b]*Harvard University Herbaria, 22 Divinity Avenue, Cambridge, MA 01238, USA*
[c]*Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA*
[d]*Agriculture and Agri-Food Canada, Wm. Saunders Building, Ottawa, Ontario K1A 0C6, Canada*

Abstract

Data curation is critical for scientific data digitization, sharing, integration, and use. This paper presents Kurator, a software package for automating data curation pipelines in the Kepler scientific workflow system. Several curation tools and services are integrated into this package as actors to enable construction of workflows to perform and document various data curation tasks. The integration of Google cloud services (e.g., Google spreadsheets), allows workflow steps to invoke human experts outside the workflow in a manner that greatly simplifies the complex data handling in distributed, multi-user curation workflows. The Kepler platform provides the modeling, execution and management ability, including a collection-oriented model of computation (COMAD), and provenance tracking and browsing for the curation package. These features not only allow workflows to be easily modeled, maintained, and evolved, but also QA/QC of curation results is facilitated through examination of provenance information recorded during workflow execution. Effectiveness of the Kurator package is demonstrated through a workflow for data curation of natural science collections.

Keywords: data curation; scientific workflows; biodiversity informatics

1. Introduction

Data curation is critical for scientific data digitization, sharing, integration and utilization. For example, in biodiversity research, on the order of a billion records of historic biological specimen data recorded in paper-based formats over the last several hundred years need to be thoroughly cleaned when they are digitized. Such data curation includes filling in missing data, correcting errors and maintaining currency. When specimen data collected from different sources that are based on different standards are compared or analyzed to answer a specific research question (such as comparing changes in species distribution over time to patterns of global environmental change),

[1] Corresponding author.
E-mail address: ludaesch@ucdavis.edu

1877-0509 © 2012 Published by Elsevier Ltd.
doi:10.1016/j.procs.2012.04.177

normalization and evaluation for fitness of use are necessary data QC steps. Thus, the challenge to make this critical scientific data resource openly available represents an enormous overhead to the natural science collection curators, which can only be realized through innovation, efficiency, and standardization.

Once the specimen data is digitized, data curation typically involves a workflow derived using multiple tools and/or services. Workflows vary based on the curation purpose or the content of the dataset(s). In this domain, examples may include a curator cleaning data recently input into a collection database by students; researchers submitting datasets they have aggregated which they wish to clean and enhance before analysis; and researchers submitting datasets to assess whether they are fit for addressing a particular hypothesis. Typically, such curation workflows are built manually. Each curation step is implemented as an individual program using one or more programming languages to access the corresponding services and tools. Input and output data at each step are often organized as files in specific formats. Necessary data format conversions between input and output data files of adjacent curatorial steps need to be handled by the data curators. Other advanced features, e.g., automatic data provenance tracking or fault-tolerance measures, usually need to be implemented from scratch or may not be available at all. To execute a curation workflow, the curator either manually invokes each curation step in turn or automatically runs through the workflow by using a custom script. Such "one-of-a-kind" curation workflows are usually hard to document, maintain, extend, and share.

Compared to this manual approach, Kepler scientific workflow technology provides an alternative solution, which helps curators automate and document data curation pipelines via workflow construction, scheduling and management. We show the effectiveness of this approach through *Kurator*, our Kepler data curation package.

2. Kurator: A Kepler Data Curation Package

2.1. Composition of Curation Tools and Services

The package consists of a number of actors for curation tasks and sample biodiversity curation workflows. The categories of actors in the package are summarized in Table 1.

Table 1. Categories of actors in the Kurator package

Curation Operation	These actors provide common, domain-independent functions useful in data curation, including data clustering, data fusion, and boundary inspection.
Authorization	Based on OAuth (Open Authorization) or Google AuthSub protocol, these actors enable authentication/authorization for access to Google services, such as Spreadsheet, GMail, etc.
Data Sharing	These actors enable data sharing through various operations on a Google spreadsheet, including spreadsheet copy, share, import, export, query, etc.
Biodiversity-Specific Curation Service	These actors provide curation services specifically useful in the biodiversity domain to show or correct data quality issues, including Google Maps, controlled vocabularies, georeferencing services, scientific name resolution services, etc.
Utility Service	These actors provide utility services, e.g., notification emails or SMS text messages, data import or querying from CSV files, etc.

Diverse services and tools can be conveniently integrated into a workflow through actors, supporting various aspects of data curation. Figure 1 shows an example workflow for specimen data curation (a demonstration video is also available [1]).

Data Cleaning Services. The example workflow reads in a CSV-formatted specimen dataset with "quality issues." Multiple biodiversity domain-specific services (e.g., GEOLocate [2], the International Plant Names Index [3], Global Names Index [4], phenological[2] analysis with authoritative data from Flora of North America [5]) are accessed to identify and correct georeferencing, scientific name discrepancies, and temporal occurrence errors. For more accurate curation, or to notify the interested parties about the discovered quality issues, the workflow also communicates with an instance of a FilteredPush network [6] in which multiple specimen data providers are

[2] Here we refer to plant flowering and fruiting periods.

connected and through which the quality information is propagated. Finally, the dataset is visualized through Google Maps to help curators visualize the result and potentially highlight other unsolved quality issues.

Collaborative, User-Interactive Features. Some quality issues may not be solved automatically based on the available services and instead may require intervention by human domain experts. The example workflow shows how participation of multiple human curators can be interwoven with automatic curation steps to yield a semi-automatic, collaborative workflow that can address more complex quality problems. To support such human-interactive features, Google cloud services are integrated for data sharing, display and editing: the to-be-curated data is imported into a Google spreadsheet and shared with a set of curators, who are notified about their curation assignment by email or SMS text message. The set of curators, their email addresses, and their areas of expertise are themselves defined in a master ("curation dispatcher") spreadsheet. After receiving their curation tasks, curators can evaluate and edit the assigned data records in their online spreadsheets, providing the corrected data and any further comments or justifications. The Kepler workflow regularly polls the status of the online spreadsheets until the curators submit their revisions, resulting in updates to one or more workflow instances. When the required number of results have been received, a final curation summary report is generated. It consists of a synthesis of proposed changes, remaining problems, and expectations about proposed actions to be taken on the original dataset. This report again is imported into a Google summary spreadsheet for final human review.

As demonstrated by the Kurator prototype, Google cloud services, especially spreadsheets, can greatly simplify complex data handling in distributed, multi-user curation workflows. In particular, the Kurator package provides a new, innovative manner to *invoke human experts as actors in a workflow*, using automatic email requests, followed by asynchronous result submission by the human experts back into the running workflow via cloud-based spreadsheets (for further illustration, see the demonstration video [1]). Through the use of the Kepler Kurator package executable tasks and steps that users would normally perform outside of Kepler can now also be orchestrated by Kepler programmatically. Kepler actors in the Kurator package are reusable to assemble workflows with different curation purposes or for datasets with different features or quality problems. Moreover, the actors are designed to be highly configurable, e.g., the dataset-clustering actor can be configured with a specific clustering algorithm and the georeference-validation actor can be configured to use different georeferencing services [2,7].

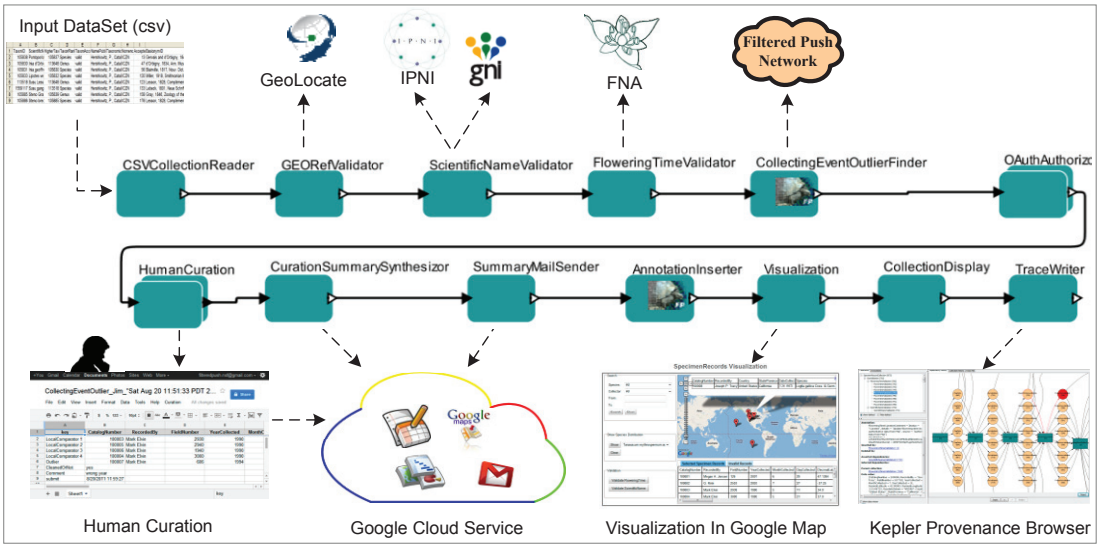

Fig. 1. Specimen data curation example workflow

2.2. Example Curation Workflow and Collection-Oriented Modeling and Design

Dataflow-oriented scientific workflow technologies, including Kepler workflows, tend to involve "complex wiring" of actors for process control and often require adaptors or "shims" [8] for data format transformations, data

selection, etc. As a consequence, the original conceptual model of the scientific data processing or analysis pipeline is sometimes hard to recognize in the complex wiring. In addition, workflows with complex wiring and many shims tend to be "brittle" and are not very resilient to changes in data formats or workflow functions.

The Kurator package uses a conveyor-belt inspired model of computation, i.e., the Kepler COMAD[3] director, to avoid or at least mitigate these problems [9,10,11,12]. The nested data collection, which is the typical structure of the scientific data, is flattened into a physical token stream with delimiters similar to an XML SAX-stream. Like workers on an assembly line, Kurator software actors are usually arranged in series and configured to operate on different parts of the data stream in a pipeline manner.

Benefiting from the COMAD model of computation, the linear structure of the curation workflow more clearly reflects the conceptual design of the curation pipeline and avoids complex actor wiring (Fig. 1). Such a workflow is thus easier to construct, evolve, and maintain. In the data streams that flow through the workflow, each data collection and each individual data element may have different tags attached. The paradigm of *tag-based processing* in workflows is facilitated by COMAD, and allows tags to be used, e.g., as flags to signal the status of data collections or elements. Downstream actors can then be configured easily to filter the dataset for different curation operations. Moreover, multiple actors can work on different parts of the data stream simultaneously, which can be exploited to greatly improve parallel performance [10,11].

2.3. Transparency and Quality Control through Data Provenance

In the Kepler/COMAD model, detailed provenance records of the workflow execution can be recorded, including data lineage, actor invocation dependencies, and the temporal evolution (processing history) of data collections. In Kepler/COMAD, provenance information[4] is embedded in the data stream and can be exported to a trace file easily by adding a TraceWriter actor at the end of the workflow pipeline (Fig. 1). Any interested parties, e.g., curators or data users, can then assess the quality and credibility of the assembled, curated data by examining the data provenance in the Kepler provenance browser [13], displaying curatorial changes made by each curator or software agent. The operations made in the curation pipeline are depicted as a graph, which can be traversed and queried easily in the provenance browser.

3. Conclusion

Kurator, a Kepler software package, facilitates automation of data curation pipelines via workflow construction, scheduling, and management. Various curation tools and services, including domain-specific services and generic Google cloud services, are integrated into this package to automate or semi-automate data curation workflows, including ones that require the participation (and implicit collaboration) of distributed human actors, i.e., curators and other domain experts. This use of Kepler to invoke actions on human actors outside of the fully automated computational workflow represents a novel expansion of workflow processes, usually seen only in business workflows, but not in scientific workflows. The use of the COMAD director, which implements a model of computation based on an assembly-line metaphor, often results in workflow designs that more closely resemble the high-level, conceptual models of workflow creators and users, compared to conventional designs [9,10]. In addition, the underlying dataflow process network model can be used to improve the performance by scheduling the actors' work in a streamlined, parallel manner. Data lineage and other fine-grained provenance information is recorded during workflow execution and can be easily examined, e.g., to understand and reproduce curation results and to assess curation quality. The Kurator package and system presented here is a working prototype. Through developing Kurator, we appreciated the generality of the Kepler system, which allowed us to extend a tool designed for documenting and executing scientific analyses to a domain requiring digital object curation processes. In doing so, we realized that the current quite technical user interface can present an obstacle to adopting and using the Kurator package by domain data curators. In future work then, we plan to develop new techniques to deal with a number of remaining challenges, including the user interface, the handling of very large datasets, and the challenges arising from the nature of long-running workflows with humans in the loop (curation tasks are assigned by email to

[3] Collection-Oriented Modeling & Design [9]

[4] The COMAD provenance model captures more fine-grained provenance than the default Kepler provenance recorder, i.e., not all output tokens depend on all input tokens read previously, and a smaller set of actual dependencies is used instead.

humans, and collected through online spreadsheets, possibly weeks later). A promising starting point seems to be the fault-tolerance and recovery/resume capabilities described in [14].

Acknowledgements

This work was supported in part by NSF awards DBI-0960535, OCI-0722079, and DOE DE-FC02-07ER25811.

References

1. L. Dou, Building specimen-data curation pipelines using Kepler workflow technology in a Filtered-Push network, 26th Annual SPHNC Meeting, San Francisco, May 26, 2011, http://www.youtube.com/watch?v=DEkPbvLsud0

2. Rios, N. E. & Bart, H. L. (2010). GEOLocate (Version 3.22) [Computer software]. Belle Chasse, LA: Tulane University Museum of Natural History.

3. The International Plant Names Index (2008). Published on the Internet http://www.ipni.org. [accessed February 2012]

4. Global Names Index, Published on the Internet http://gni.globalnames.org/

5. Flora of North America, Published on the Internet http://fna.huh.harvard.edu/

6. Z. Wang, H. Dong, M. Kelly, J.A. Macklin, P.J. Morris, and R.A. Morris, Filtered-Push: A Map-Reduce Platform for Collaborative Taxonomic Data Management, Computer Science and Information Engineering, 2009 WRI World Congress on, pp.3:731-735, 2009

7. BioGeomancer, Published on the Internet http://biogeomancer.org/

8. C. Lin, S. Lu, X. Fei, D. Pai, and J. Hua. A task abstraction and mapping approach to the shimming problem in scientific workflows. IEEE Intl. Conf. on Services Computing, pp. 284-291, 2009.

9. T. McPhillips, S. Bowers, D. Zinn, and B. Ludäscher. Scientific workflow design for mere mortals, Future Generation Computer Systems, pp. 25(5):541–551, 2009.

10. D. Zinn, S. Bowers, T. M. McPhillips, and B. Ludäscher. Scientific workflow design with data assembly lines, 4th Workshop on Workflows in Support of Large-Scale Science, 2009.

11. D. Zinn, S. Bowers, and B. Ludäscher. XML-based computation for scientific workflows, 26th IEEE International Conference on Data Engineering, pp.812–815, 2010.

12. L. Dou, D. Zinn, T. McPhillips, S. Köhler, S. Riddle, S. Bowser, and B. Ludäscher, Scientific workflow design 2.0: Demonstrating streaming data collections in Kepler, 27th IEEE International Conference on Data Engineering, pp.1296-1299, 2011.

13. M.K. Anand, S. Bowers, and B. Ludäscher, Provenance browser: Displaying and querying scientific workflow provenance graphs, 26th IEEE International Conference on Data Engineering, pp.1201-1204, 2010.

14. S. Köhler, T. McPhillips, S. Riddle, D. Zinn, and B. Ludäscher, Improving Workflow Fault Tolerance through Provenance-based Recovery, 23rd Scientific and Statistical Database Management Conference, pp.207-224, 2011.