Journal of World-Historical Information
JWHI

RESEARCH REPORT

Evgeny Karataev and Vladimir Zadorozhny

Col*Fusion:
A global-scale information integration infrastructure based on crowdsourcing

The ability to share data is of a great importance, particularly to ensure its reuse (see “characterizing reuse” in [1]). While there are numerous data sets available from various groups worldwide, the existing data sources are principally oriented toward regional comparative efforts rather than global applications. They vary widely both in content and format. Such data sources cannot easily be integrated and maintained by small groups of developers. In this report, we describe *Col*Fusion (Collective data Fusion)*—an advanced infrastructure for systematic accumulation and utilization of global heterogeneous data based on crowdsourcing.

What Col*Fusion is

The vision. Imagine a world in which all datasets are publicly available (with appropriate permissions and licensing) and access points to those datasets are not spread over hundreds of different digital libraries and repositories, government and personal websites, etc. The process of data sharing could be as simple as visiting a web site and doing a couple of mouse clicks with no time-consuming data preparation and transformation to fit

Volume 2-3, No. 1 (2014-15) | ISSN 2169-0812 (online)
DOI 10.5195/jwhi.2015.25 | http://jwhi.pitt.edu

strict format requirements. The data could also be shared automatically from any data manipulation software.

Imagine being credited and recognized for the data you share and seeing your datasets used and/or evolving over time. The datasets you share are not getting lost among other datasets but, instead, they are automatically linked with other datasets (even from other disciplines) while preserving data autonomy. The data linkage could bring you new interdisciplinary research questions and new collaborations that you did not expect at the time you created your data.

Instead of searching for locations of useful datasets, you just look for the data you are interested in—for instance, the variables that you want to analyze. Such a search does not merely result in a list of links to the locations of potentially relevant datasets: instead, the result includes the data items that you are interested in. Moreover, if the data that you need are originally located in a number of distinct datasets, the resulting dataset will automatically integrate all those datasets.

Any dataset you look at will also include comprehensive metadata—both on the dataset level (e.g., title, description, authors, etc.) and the variable level (a short description explaining what each variable stores, the data type, measuring unit, etc.). The dataset will also be provided with provenance information that includes any actions performed on the dataset since it was created. In the case of an integrated dataset, provenance information will describe how the integration was performed (e.g. which datasets were used, which variables were used, any transformations that were applied, etc.). If the datasets were used in any published work, you will be able to obtain and review a list of relevant papers. You will be able immediately to explore and understand the dataset by looking at data visualizations (graphs, maps, etc.) created and shared by other users. You will also be able to reuse the results of previous research, such as code and statistical models for analyzing the datasets.

Imagine that you could join other researchers currently working on datasets that interest you and communicate directly with them while working on those datasets. Moreover, you could run complex data analysis algorithms and write your own data analysis or visualization program collectively, without the need to download the datasets to your local machine. Instead, you would utilize the power of the high-performance cloud computing. After that you could easily share your analysis with the research community. For complex tasks that are out of your competence, you could also hire (for money or other rewards) domain experts, programmers, and data analysts.

You could write papers in which all charts are interactive, produced via simple user interfaces. Those papers would be published online, allowing other researchers to try different parameters to better understand your work. In such papers, the data and analytical code become a major part of the paper, not just something nice to have. This would allow other researchers to validate your studies.

Imagine that all features explained above are implemented in an east-to-use web application that requires neither complicated installation nor considerable learning efforts. All of the above describes what we see Col*Fusion to be. Below we elaborate on the current state and the future of our work.

Current State. Col*Fusion is in active development (currently it includes about 60K lines of code). Figure 1 shows the major components of Col*Fusion architecture (gray boxes represent unfinished modules and functionality). Col*Fusion is designed in a modular fashion that makes it easy to replace specific modules if needed as well as to distribute Col*Fusion execution among a cluster of machines. Col*Fusion hardware and software architecture is designed to enable effective data integration and analysis through crowdsourcing at the *interactive rates* that people expect of web-based resources.

Col*Fusion architecture consists of four internal components: (1) the *Access Layer* component that provides several interfaces so that users and third party tools can interact with Col*Fusion, (2) the *Col*Fusion Core* component that is responsible for the business logic and metadata management, (3) the *Distributed Data Processing* component that handles large scale data processing on a distributed set of commodity machines, (4) the *Replicated Distributed Data Storage* component that is responsible for storing all datasets. In addition there are two external components: (1) the *Dataverse Network* [2][3] project developed at Harvard that is used as data archive, and (2) the *Pittsburgh Supercomputing Center* [4] facilities used as the large scale high-performance computing resource.

At the moment, the data submission module allows users to submit data from Excel, CSV, and database dump files. During data submission, Col*Fusion tries to collect as much metadata as possible on both the dataset level and the variable level. The metadata are either retrieved from the data file (e.g., variables names, data types, etc.) or entered by the user (e.g., title, description, tags, category, variables format and measuring units, etc.).

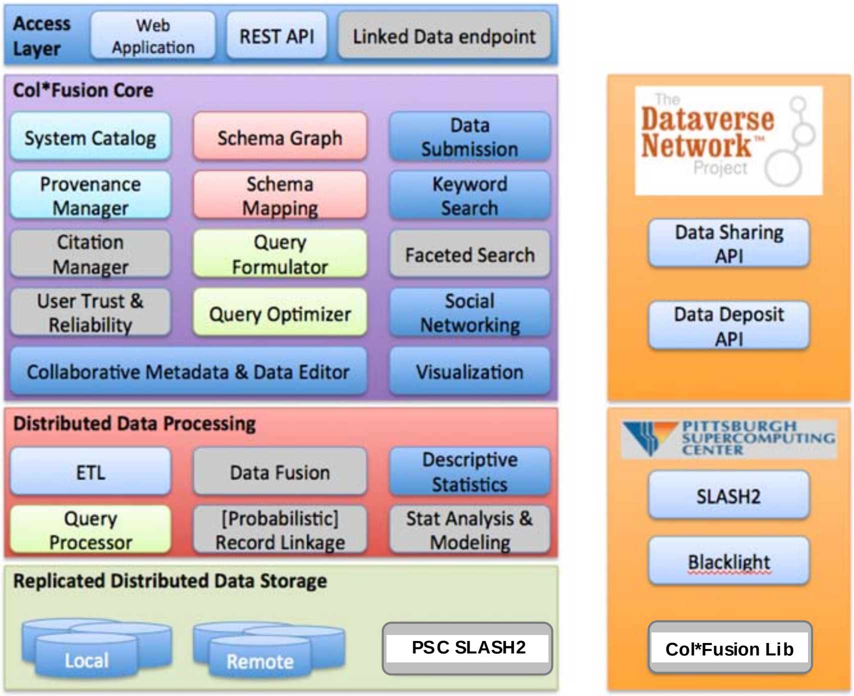

Figure 1. Col*Fusion Architecture (gray boxes represent unfinished modules/functionality).

For each submitted dataset, Col*Fusion creates a dedicated, identifiable page (we call it "Story page") on which users can view the metadata and data in a tabular, paged format. Both metadata and data can be collectively edited any time after submission by registered Col*Fusion users. Data editing is supported by an integrated OpenRefine [5] editor that provides basic and advanced cell and column edit and transformation functionalities. The Provenance Manager keeps track of all changes and performs automatic versioning that makes it possible to roll back any undesired changes.

Once the user confirms data submission, Col*Fusion performs information linkage, which includes schema mapping and discovery of relationships between the newly submitted dataset and existing datasets in the Col*Fusion repository. Currently, Col*Fusion establishes a relationship between datasets based on linguistic similarity over variables metadata in those datasets. Col*Fusion users can provide feedback on automatically generated relationships in terms of confidence values that reflect users' belief that relationships hold. Also, users can manually create a relationship if Col*Fusion fails to identify one.

Col*Fusion provides a keyword search functionality that is quite distinctive. Keyword search functionality in existing data repositories only searches through dataset-level metadata and returns a list of datasets that might contain data that the user is interested in. The result of a Col*Fusion search is not just a list of datasets that might have the data of interest but rather a *merged* dataset—or a list of merged datasets if there are several possible paths to answer the query.

Current and future work. As we mentioned above, we are actively working on Col*Fusion implementation that includes improving stability and performance as well as adding new features towards the vision presented in section 2.1. Here are some of the current and future tasks. First of all, to utilize the functionality of established data repositories—such as data archival and preservation, citation, and metadata export in Dublin Core [6][7] and DDI [8][9] formats—we are working towards integration of Col*Fusion with the Dataverse Network over the Dataverse Sharing and Deposit APIs. This will allow Col*Fusion to fetch/deposit datasets directly from/to Dataverse Network.

Since Col*Fusion actually processes the submitted data files, with proper interface it can be utilized as a cloud data analytical platform, where the researcher would be able to run complex data visualization and analysis tasks. We plan to integrate IPython Notebook [10] into Col*Fusion. IPython Notebok (currently evolving into Jupyter project [11]) is "… web-based interactive computational environment where you can combine code execution, text, mathematics, plots and rich media into a single document." Thus, in the future, in addition to the integrated data repository, we envision Col*Fusion to be a virtual collaborative research environment where researchers can communicate their ideas, run analysis on the integrated data, build and share interactive visualizations, and even write and publish online interactive research papers with data and analytical code integrated into them.

Current implementation of the keyword search allows users to search for data transparently throughout all submitted datasets without prior knowledge of any query language or the relationships between the datasets. To provide support for more complex queries without forcing users to learn any particular query language, we plan to implement a faceted search. Finally, to connect integrated data from Col*Fusion to the Linked Data cloud [12], we plan to implement a Col*Fusion Linked Data endpoint that will allow Col*Fusion data to be presented and queried according to linked data principles.

Conclusion

More and more data become available every day; at the same time a lot of data that are never reused as datasets become scattered in silos around the globe. Many data-curation and archival tools have been developed over the last twenty years (a recent one is Tamr [13][14]), but most of the enterprise data integration approaches focus on data integration on a customer basis. For example, an enterprise that has a number of operational databases might face the task of integrating them. A much bigger goal is to create a global data space where any data instance can be reached. The Semantic Web with Linked Data principles has this goal: "The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. . . . With linked data, when you have some of it, you can find other, related, data. " [15]. However simply following the Linked Data principles to publish research data would not ensure the reusability of that data, and for many reasons: data provenance, quality, credit, attribution and reproducibility. For example, publishing data out of context would fail either to reflect the research methodology or to respect the rights and reputation of the researcher [16].

Therefore, we believe that by combining sophisticated algorithms utilized in enterprise data integration systems and various data management techniques and tools with the power of crowdsourcing, Col*Fusion can fully realize the vision set forth at the beginning of this report.

NOTES

[1] Etal, B. et al. 2013. "Why linked data is not enough for scientists." *Future Generation Computer Systems*. 29, 2 (Feb. 2013), 599–611.

[2] Crosas, M. 2011. "The Dataverse Network®: an open-source application for sharing, discovering and preserving data." *D-Lib Magazine*. (2011).

[3] King, G. 2007. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." *Sociological Methods & Research*. 36, 2 (Nov. 2007), 173–199.

[4] Pittsburgh Supercomputing Center: http://www.psc.edu/. Accessed: 2015-07-17.

[5] OpenRefine: http://openrefine.org/. Accessed: 2015-07-16.

[6] Dublin Core: http://dublincore.org/. Accessed: 2015-03-29.

[7] Weibel, S., Kunze, J., Lagoze, C. and Wolf, M. 1998. "Dublin core metadata for resource discovery." *Internet Engineering Task Force RFC*. 2413, 222 (1998), 132.

[8] Data Documentation Initiative (DDI): http://www.ddialliance.org/. Accessed: 2015-03-29.

[9] Ryssevik, J. 2001. "The Data Documentation Initiative (DDI) metadata specification." Ann Arbor, MI: Data Documentation Alliance. Retrieved from http://www. ddialliance.org/sites/default/files/ryssevik_0. pdf (2001).

[10] Perez, F. and Granger, B.E. 2007. "IPython: a system for interactive scientific computing". *Computing in Science & Engineering*. 9, 3 (2007), 21–29.

[11] Jupyter: https://jupyter.org/. Accessed: 2015-07-14.

[12] Bizer, C., Heath, T. and Berners-Lee, T. 2009. "Linked data—the story so far." *International journal on Semantic Web and Information Systems*. 5, 3 (2009), 1–22.

[13] Tamr: http://www.tamr.com/. Accessed: 2015-07-12.

[14] Stonebraker, M., Ilyas, I.F., Zdonik, S., Beskales, G. and Pagan, A. 2013. "Data Curation at Scale : The Data Tamer System." *CIDR* (2013).

[15] Linked Data - Design Issues: http://www.w3.org/DesignIssues/LinkedData.html. Accessed: 2014-04-09.

[16] Etal, B. et al. 2013. "Why linked data is not enough for scientists." *Future Generation Computer Systems*. 29, 2 (Feb. 2013), 599–611.

This journal is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.