DIBBs Brown Dog – The Need for and Challenges of a Science Driven Data Transformation Service

Kenton McHenry, Shannon Bradley, Michael Dietze, Praveen Kumar, Jong Lee, Richard Marciano, Luigi Marini, Jerome McDonough, Barbara Minsker, Art Schmidt, Bill Sullivan

ACI-1261582

With growing and diverse collections of data becoming part of modern scientific workflows, many research projects today begin with a process of data wrangling, i.e. finding, manipulating, indexing, cleaning, and bringing together needed datasets. Brown Dog aims to alleviate much of the overhead and heterogeneity in the processes involved in this step which tends to otherwise hinder scientific progress and reproducibility. Through a REST API Brown Dog provides data transformations such as format conversions and content based extractions as a service which supports diverse usage through various clients and programming languages. Further, Brown Dog provides a venue to access and preserve data transformation tools, track provenance, track information loss, manage data movement, and process jobs in a scalable manner across a diverse set of computational resources. Overall, Brown Dog provides a low-level data infrastructure to interface with digital data contents and through its capabilities move software to being more agnostic to the format/structure of data, enabling the scientific community to focus more on their research, less on data wrangling, and allow researchers to more easily access datasets that would otherwise be inaccessible.

Rooted within recent academic software activities over the past couple decades, within countless projects across every scientific domain, the notion of Brown Dog and an extensible scientific data transformation service further aims at reducing the cost, redundancy, and many one-off efforts done within science surrounding ingestion and cleanup of data. Countless tools are built, costing significant amounts of time and money, only to be lost and repeated in other efforts. By taking a step back from the individual science activities, observing the commonalities, and building a service at a sufficiently low level, we put into place a resource that all communities can then leverage to reduce the time, effort, and variability in how they work with data. As a web service, which can be leveraged by most modern programming languages and tools, Brown Dog serves as a lowest common denominator of what's needed in most circumstances within tools across domains that would need to transform data as part of their workflow.

A key challenge within modern scientific software solutions, however, is the high variability in programming skills amongst the scientists and students across the various scientific domains. Thus, things like a web service with a REST API, while widely accessible and leverageable across programming languages and tools, can be difficult for users that only have a moderate level of programming ability. Towards dealing with this broad diversity in skills Brown Dog developers have worked closely with researchers in various domains toward prototyping interfaces within tools familiar to the community. Examples of this include plugins for ArcGIS for geospatial data manipulation, QGIS an open source alternative for geospatial data manipulation, Microsoft Excel for numerical data manipulation, and libraries within languages such as Matlab, R, and Python. More generally we have also built broadly usable interfaces such as a web bookmarklet that modifies the search function on web pages, allowing one to search the contents of files on a page, and adds menus to URLs allowing one to download files in a variety of alternative formats. Other interfaces include a Microsoft Windows plugin, a Unix command line interface, and a new web interface allowing a user to easily try out the service while also being exposed to the various programmable interfaces to Brown Dog, called BDFiddle. As a web service first and foremost Brown Dog naturally allows for such diverse interfaces and client applications.

As programing is becoming a more and more crucial part of scientific research, efforts such as the Software Carpentry and Data Carpentry workshops have worked to help expose these valuable skills to more and more students across non-computer science fields. The Brown Dog team has already begun to follow suit, having had its first tutorial session at XSEDE 2016. We plan to continue to engage directly with scientists, prototype interfaces within community tools, provide regular tutorials, and engage with the Software Carpentry movement towards enabling greater numbers in the scientific community to develop their applications around such resources.