Description

The information age has made it trivial for anyone to create and then share vast amounts of digital data. This includes unstructured collections made of data such as images, video, and audio to collections of born digital content made up of data such as documents and spreadsheets. While the creation and sharing of content has been made easy, its inverse, the ability to search and use the contents of digital data, has been made exponentially more difficult. In the physical analogue librarians have used the process of curation to standardize the format by which information is stored and diligently index holdings with metadata to allow both current and future generations to find information. Digitally this does not happen as that curation overhead is an unwelcomed bottleneck to the creation of more data. Though popular services such as modern search engines give the illusion that this is being done, this is largely over the portion of digital data that is text based and/or containing text metadata. Unstructured collections and contents trapped behind difficult to read file formats, however, make up a significant part of our collective digital data assets and are largely not accessible.

Science today not only uses but relies on software and digital content. It is well known that science is not only responsible for a significant amount of our digital data holdings but that also much of this is un-curated data, what the scientific community currently refer to as “long-tail” data. As such contemporary science, which relies on digital data and software, software which evolves and disappears quickly as underlying technology changes, is entering a realm where scientific results are no-longer easily reproducible and as such in essence no longer a science as science hinges on the fact that a documented procedure will result in the same result each time.

Brown Dog’s goal is to construct a Data Transformation Service (DTS) that will play a role similar to a Domain Name Service (DNS), existing as part of the Internet’s backbone for the purpose of making the contents of un-curated data collections more accessible. The DTS will support data conversions by building off of a technology called a Polyglot Software Server which replaces an applications native interface with a uniform interface that can be easily programmed against. Using Software Servers the DTS will chain together open/save operations within software applications for the purpose of seamlessly transforming unreadable files to readable ones, making future browsers and applications more agnostic to file formats. Second, the DTS will support metadata extractions by building off of the Clowder framework for data management, sharing, curation, and publication. Using Clowder the DTS will be able to serve as an active repository for both housing and utilizing content analysis software within the community for the purpose of indexing and automatically assigning metadata to un-curated collections. Supporting these two types of transformations will allow the DTS to act as a DNS for data, translating in-accessible un-curated data into information in a manner that is provenance preserving so as to ensure reproducible science and enable new science over our vast collections of un-curated digital data. The intellectual merit of this work lies in the proposed solution which does not attempt to construct a single piece of software that magically understands all data, but instead aims at utilizing and interconnecting every possible source of automatable help already in existence (i.e. software, libraries, code) in an extensible, robust, and scalable manner to create a service that can deal with as much of this data as possible. This proverbial “super mutt” of software, or Brown Dog DTS, will serve as a low level infrastructure to data enabling a new era of science and applications which will not only make use of, but rely upon un-curated data sources. The broader impact of this work is in its potential to serve not just the scientific community but the general public, as a DNS for data, moving civilization towards an era where a user’s access to data is not limited by a file’s format or un-curated collections. Three use cases spanning geoscience, biology, engineering, and social science are proposed to both drive and demonstrate the novel science that can be obtained with this underlying cyberinfrastructure.