# Brown Dog Introduction for New Graduate Students

As part of the National Science Foundation's data efforts Brown Dog aims to provide and preserve long term access to data within collections of unstructured and uncurated files. Two services will be developed. The Data Access Proxy (DAP), an extensible and distributed file format conversion service, aims at making the web agnostic to the file formats which information is stored in and allowing users to more easily gain access to the contents of files. The Data Tilling Service (DTS) will serve as an active repository for analysis tools and provide a service by which to automatically extract metadata and content based signatures from files over the web. Together these two services stand as a building block between raw data collections and applications that would provide search capabilities of such collections and/or organize, relate, curate, or otherwise use their contents.

As a graduate student on the project your aim is first and foremost to carry out the research as directed by your adviser towards addressing a specific scientific question within your field which requires examining collections of unstructured and/or uncurated data. Think of unstructured data as data types that do not have a pre-defined data models or are not organized in a pre-defined manner. Unstructed data can be text based but can also involve sensor data or data that quantifies some physical object or phenomanon (e.g. images, video, audio, 3d models, etc.). Such data is typically difficult to understand using traditional computer programs. Images are a good example of this. To a computer images are nothing more than an array of numbers representing pixel intensities or colors. Though images are extremely informative to us as human beings, for a computer to make any use of them some form of pre-processing must be run. An example would be to use computer vision to recognize faces within the image and then spit out their locations as numerical values and a textural tag identifying these areas as faces. With information such as this a computer is then more readily able to carry out a search or other process involving the contents of such data. With regards to uncurated data think of a dump of some random hard drive. Without meaningful file names and a meaningful directory structure it will be difficult to find information without examining each and every file. File formats, in particular old and/or proprietary file formats, hinder the situation further by making it difficult to open a given file without the needed software to open it installed on your machine. Metadata is another way of providing insight as to the contents of a file. Consider a document tagged with keywords "paper, large dynamic groups" indicating a paper submission for a social science study looking into the behavior of large groups of people. Curated data is data that has been stored and diligently named, organized, and tagged so that others, both today and long in the future, can utilize the data. Uncurated data on the other hand doesn't have much of this and is essentially a big mess for others to go through. A significant amount of digital data, if not most, is uncurated. In the scientific world this is sometimes referred to as "long tail" data, suggesting this is linked with the tail of the distribution of project sizes, with the vast majority of smaller projects not having the resources to properly manage the data they produce. The bottom line is that curation is a cumbersome process and creating new data is both faster and more rewarding, at least in the short term, than going back and organizing old data. As science hinges on reproducibility and building on past results, however, these problems must be addressed.

As part of your work you will do one or both of the following. You will develop your own analysis tools for the specific collections you will be looking at, towards the specific scientific questions you are addressing. These tools will be built such that they can be included into the DTS for preservation, reproducibility of your results, and reuse by other in the scientific community (possibly for things completely different from that which you used it). You will build novel applications to utilize the information within the specific collections you will be looking at that utilize tools already within the DAP and/or DTS services. NCSA software developers will aid you in including needed functionality specific to the collections you are looking at. As part of this you will provide information with regards to where the data collection resides, the file formats it contains, the file types it contains, and the information that needs to be extracted from this data.

Before getting started please review the relevant sections of the Brown Dog reading list, containing papers provided by your adviser for your specific area. The reading list can be found on the project wiki:

https://opensource.ncsa.illinois.edu/confluence/display/BD

under "Information and Resources" along with other useful information such as the location of the projects source code repositories, issue tracker, the mailing lists for the project, names and emails of the project team, links to tutorial videos for the software we will be building off of, and the "Use Case Guidelines" outlining how your work should integrate with the overall project. Other than that please get to know and get comfortable communicating with the Brown Dog developers at NCSA and UNC. We expect to communicate frequently and we look forward to working with you.

*Kenton McHenry*

*Brown Dog PI*