Definitions

Brown Dog Data Transformation Service (DTS): A distributed internet service providing data transformations as a service (e.g. conversions, extractions). Clients, or users, make requests of the DTS via a REST API (querying capabilities, requesting transformations, etc) offloading the need to do these transformations locally and/or across many applications. New transformations are added to and managed by a central catalog from where they can then be deployed to various DTS instances. The DTS has thus far been utilized to support transformations with Python, R, Matlab, Java, from the Linux/Mac command line, from the Windows file manager, from ArcGIS and QGIS, from Excel, and from web browsers.

Clowder: The Brown Dog component responsible for extracting novel, often higher level, data from file contents (e.g. metadata, tags, signatures, and other derived products) in order to index, compare, and further analyze collections of data through a broadly accessible REST API. Clowder is a web based research data management system designed to support multiple research domains and the diverse data types utilized across those domains. In addition to data sharing and organizational functionality it contains major extension points for the preprocessing, processing, previewing, and publication of data. When new data is added to the system, whether it is via the web front-end or through the REST API, preprocessing serving as a form of autocuration is off-loaded to cloud based extraction services that analyze the data's contents to extract appropriate data and metadata. These extractors triggered based on the type of the data analyze the contents of the data to tag it (e.g. found flood basins in images, trees in LIDAR) and/or create lightweight web-accessible previews of large files (e.g. an image pyramid) allowing users to examine and compare the contents of one or more datasets. Complimented by a number of features supporting community based social curation, this combined raw and derived metadata is presented to the user in the Clowder web interface and utilized to navigate stored collections.

Data Conversion: A transformation on digital data that largely preserves the entirety of the data. An example in the case of Brown Dog would be a transformation of a file in one 3D file format to another 3D file format. As file formats typically vary slightly, and the transformations themselves can be imperfect, variations can occur in the form of information loss. However, the intent is for the resulting data to be as intact as possible. Conversions allow one to access data more easily given that the original format is not understood or difficult to work with. This is analogous to translating languages.

Data Extraction: A transformation that creates new data from the given data. An example in the case of Brown Dog would be the execution of analysis code on an image file's contents to determine if a particular species of plant is present. We utilize extraction to automatically generate metadata and/or signatures from a file's contents and provide users with means of finding, relating, and utilizing data that may be difficult otherwise.

Metadata: Simply data about data (e.g. tags or keywords).

Polyglot: The Brown Dog component responsible for file format conversions. Utilizing Software Servers and Daffodil, Polyglot is a highly distributed and extensible service which brings together and manages conversion capabilities, both the needed computation and data movement, from other software, tools, and services under a broadly accessible REST API.

pyClowder2: Python utility library simplifying the process of adding an analysis tool as a an extractor, wrapping most interactions with Clowder as python functions.

Software Server: Light weight web utility used to add a REST API onto arbitrary software and tools. Component within the Polyglot framework/repository used in the creation of converters**.**

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.

Unstructured Data: Data that does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured data can be text based but can also involve sensor data or data that quantifies some physical object or phenomenon (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.
