BROWN DOG

browndog.ncsa.illinois.edu

The persistent growth of "big data" presents challenges that include not only large quantities but also large varieties of digital data.

Brown Dog is a highly extensible set of services for:

- *data format conversions*
- *metadata extraction from data contents*
- *indexing uncurated collections of data*
- *preserving & using data analysis/manipulation tools*

Much of the data generated by science, social science, and the humanities is smaller, unstructured, un-curated and thus not easily shared. This vast quantity of "long-tail" data, both past and present, has the potential for great impact on future research in many disciplines.

In addition, science relies on digital data and software that evolve and disappear quickly as technology changes. We are entering a period when scientific results are no longer easily reproducible. Easily accessing legacy data and software is essential to maintaining the viability of large bodies of research. Without a consistent and uniform index over all the data, or at least associated metadata, such actions become prohibitively difficult.

Brown Dog addresses these and similar considerations through the development of a set of services for data format conversion, metadata extraction from existing data, and indexing of uncurated collections of data. Brown Dog provides the framework for an extensible suite of new and existing tools. Researchers using the scalable Brown Dog system will be able to apply the whole suite of tools to data collections in order to find and "unlock" previously inaccessible data.

The Data Access Proxy (DAP) and the Data Tilling Service (DTS), focusing on format conversions and content based analysis/extraction respectively, wrap relevant conversion and extraction operations within arbitrary software, manages their deployment in an elastic manner and manages job execution from behind a deliberately compact REST API.

A number of client libraries and applications are being constructed to further reduce the overhead of accessing the provided functionality (e.g. libraries in Javascript, Python, R, MATLAB and interfaces such as bookmarklets, browser extensions, and other standalone applications). At the heart of the two services is their extensibility, which allows potentially any library, application, or other service to be incorporated as a conversion or extraction tool. This leverages the functionality they provide and preserves the tools themselves. Both services support a variety of scripting languages to wrap tools for inclusion in Brown Dog (e.g. Python, MATLAB, R, bash, AutoHotKey, etc.)

Brown Dog aims to support data conversion and extraction/analysis needs from a broad range of communities. Current efforts focus on biology, ecology, hydrology, and social science with goals to add functionality for other geoscience communities, material sciences, and the humanities, as well as for the general public.

```
#Application name (Version)
#File types supported (e.g. document, depth, image, ...)
#Comma separated list of supported input formats
#Comma separated list of supported output formats
                                                    Describe

#Call external application and/or carry out conversion
...                                                 Convert File
```

Comments needed to include a new tool as a converter within the DAP

```
extractors.connect_message_bus(extractorName=extractorName,
                               messageType=messageType,
                               rabbitmqURL=rabbitmqURL,
                               rabbitmqExchange=rabbitmqExchange,
                               processFileFunction=process_file,
                               checkMessageFunction=check_message)
                                                    Connect

def process_file(parameters):
    global extractorName
    inputfile=parameters['inputfile']

    # call actual program
    result = subprocess.check_output(['wc', inputfile], stderr=subprocess.STDOUT)
    (lines, words, characters, filename) = result.split()   Work on File

extractors.upload_file_metadata(mdata=metadata,
parameters=parameters)
                                                    Return Metadata
```

pyClowder library calls needed to include a new Python tool as an extractor within the DTS

HOW TO INCLUDE IN DATA MANAGEMENT PLANS

http://browndog.ncsa.illinois.edu/datamanagement

The data analysis and conversions developed here will be pushed into the NSF DIBBs: Brown Dog (ACI-1261582) project as data extractors/converters within the DTS and DAP, services providing automatic data annotations/analysis and format conversions as broadly usable internet resources. Brown Dog aims to both provide services and tools to aid in the curation, accessing, and indexing of data as well as to preserve scientific software that might be leveraged for that purpose. As Brown Dog extractors/converters, the capabilities of these tools will be preserved, will take part in an ecosystem of other extraction/conversion tools, and will be leverageable by others within the scientific community, perhaps in very different fields, as well as by the general public.

This research and development has been funded through National Science Foundation Cooperative Agreement ACI-1261582.

BROWN DOG

USER CHEAT SHEET

THE DAP REST API FOR CONVERSIONS

GET	/api/conversions/outputs	Lists all output formats that can be reached
GET	/api/conversions/inputs	List all input formats that can be accepted
GET	/api/conversions/inputs/{input format}	List all output formats that can reach the specified input format
GET	/api/conversions/outputs/{output format}	List all input formats that can reach the specified output format
GET	/api/conversions/convert/{output format}/{file URL}	Convert the specified file to the requested output format
POST	/api/conversions/convert/{output format}	Convert the uploaded file to the requested output format
GET	/api/conversions/software	List all available conversion software
GET	/api/conversions/servers	List all currently available Software Servers

THE DTS REST API FOR EXTRACTIONS

GET	/api/extractions/inputs	Lists the input file format supported by currently running extractors
POST	/api/extractions/url	Uploads a file for extraction of metadata using the file's URL and returns a file id
POST	/api/extractions/file	Uploads a file for extraction of metadata and returns a file id
GET	/api/extractions/{id}/status	Checks for the status of all extractors processing the file with id
GET	/api/files/{id}/metadata	Gets tags, technical metadata, and content based signatures, and other derived products extracted for the specified file
GET	/api/extractions/extractors	List currently available extractors
GET	/api/extractions/extractors/details	List details of currently avaialable extractors
GET	/api/extractions/servers	List all currently available extractor servers

This research and development has been funded through National Science Foundation Cooperative Agreement ACI-1261582.