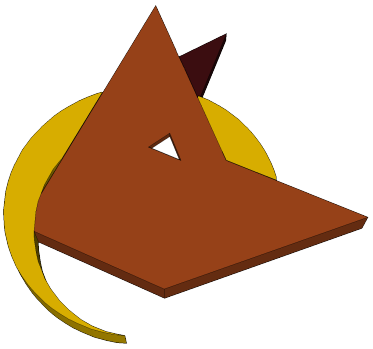
DIBBs: Brown Dog

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. DIBBs Brown Dog aims to alleviate some 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 (leveraging Polyglot) and content based extractions (leveraging Clowder) 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.

- **Documentation**
  - Introduction
  - BDFiddle
  - BD CLI
  - BD Windows Client
  - API
  - Preparing Tools
  - Adding Tools
  - Project Supported Transformations
- **Workshops & Tutorials**
  - **Brown Dog Tutorial - Beta Release**
  - NCHC Tutorial 2017
  - XSEDE Tutorial 2016
  - Early User Workshop 2015
  - Kickoff Tutorials 2013 and 2014
- **Project Information and Resources**
  - Components & Source Code
  - Data Management Plans
  - Description
  - Definitions
  - Events, Outreach, and Media
  - Project Team
  - Publications
  - Resources
- ***Discussions / Under Development***

Blog Posts

- Blog: NCSA Brown Dog and Box Skills Speed up Astronomical Research created by Shannon Bradley
  Apr 12, 2019
  BrownDog
- Blog: The Predictive Ecosystem Analyzer - PEcAn created by Kenton McHenry
  Mar 21, 2018
  BrownDog
- Blog: Using Machine Learning to Understand Public Preference Toward Landscape Design created by Kenton McHenry
  Feb 21, 2018
  BrownDog
- Blog: Brown Dog Tutorials created by Kenton McHenry
  Feb 01, 2018
  BrownDog
- Blog: What is Brown Dog Video created by Kenton McHenry
  Nov 06, 2015
  BrownDog
- Blog: IEEE Big Data created by Kenton McHenry
  Nov 06, 2015
  BrownDog
- Blog: Brown Dog Cheat Sheet created by Kenton McHenry
  Sep 16, 2015
  BrownDog
- Blog: To Be Heard and Not Seen created by Kenton McHenry



https://bd-api.ncsa.illinois.edu/

Recent Space Activity

Kenton McHenry
Events, Outreach, and Media updated Feb 19, 2020 view change
Publications updated Feb 19, 2020 view change

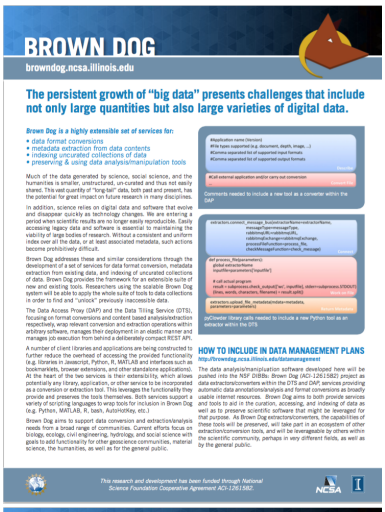
Mark Fredricksen
Transformations Catalog updated Jan 27, 2020 view change
Setting up Windows VM for ArcGIS extractors updated Jan 24, 2020 view change

Sandeep Puthanveetil Satheesan
Setting up Windows VM for ArcGIS extractors updated Jan 09, 2020 view change

Brown Dog Flyer / Cheat Sheet

Apr 27, 2015
BrownDog

- Blog: Critcal Zone created by
Kenton McHenry
Apr 25, 2015
BrownDog
- Blog: PEcAn created by
Kenton McHenry
Apr 25, 2015
BrownDog
- Blog: Green Infrastructure created by
Kenton McHenry
Apr 25, 2015
BrownDog
- Blog: Brown Dog Clients: A Command Line Interface created by
Kenton McHenry
Mar 18, 2015
BrownDog
- Blog: The Brown Dog Tools Catalog created by
Kenton McHenry
Mar 05, 2015
BrownDog
- Blog: Ecological Model Data Conversions: PEcAn created by
Kenton McHenry
Mar 03, 2015
BrownDog
- Blog: Brown Dog Clients: Chrome Extension created by
Kenton McHenry
Mar 03, 2015
BrownDog


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 Tiling 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, civil engineering, hydrology, and social science with goals to add functionality for other geoscience communities, material science, the humanities, as well as for the general public.

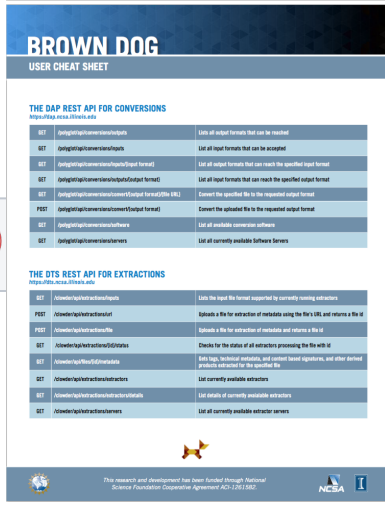
#Application name (Version)
#File types supported (e.g. document, depth, image, ...)
#Comma separated list of supported input formats.
#Comma separated list of supported output formats
Preamble
#Call external application and/or carry out conversion
Conversion
Comments needed to include a new tool as 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', '-mpwfilel', sbdm+=subprocess.STDOUT]
(lines, words, characters, filename) = result.split()
Work on data
extractors.upload_file_metadata(metadata=metadata, parameters=parameters)
Upload results
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/manipulation software 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 extractors/conversion tools, and will be leveragable 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
https://dap.ncsa.illinois.edu

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

THE DTS REST API FOR EXTRACTIONS
https://dts.ncsa.illinois.edu

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

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