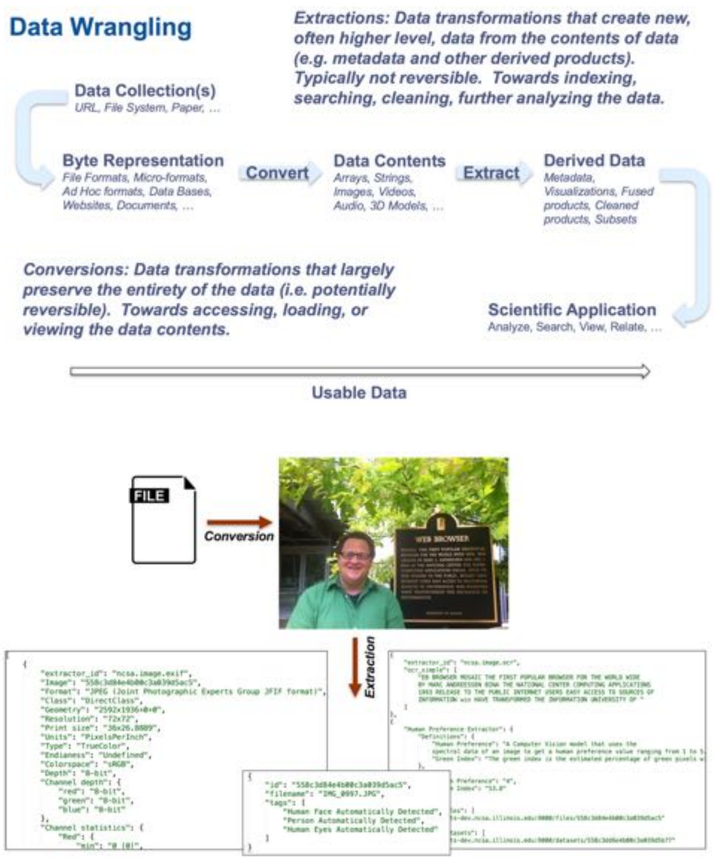
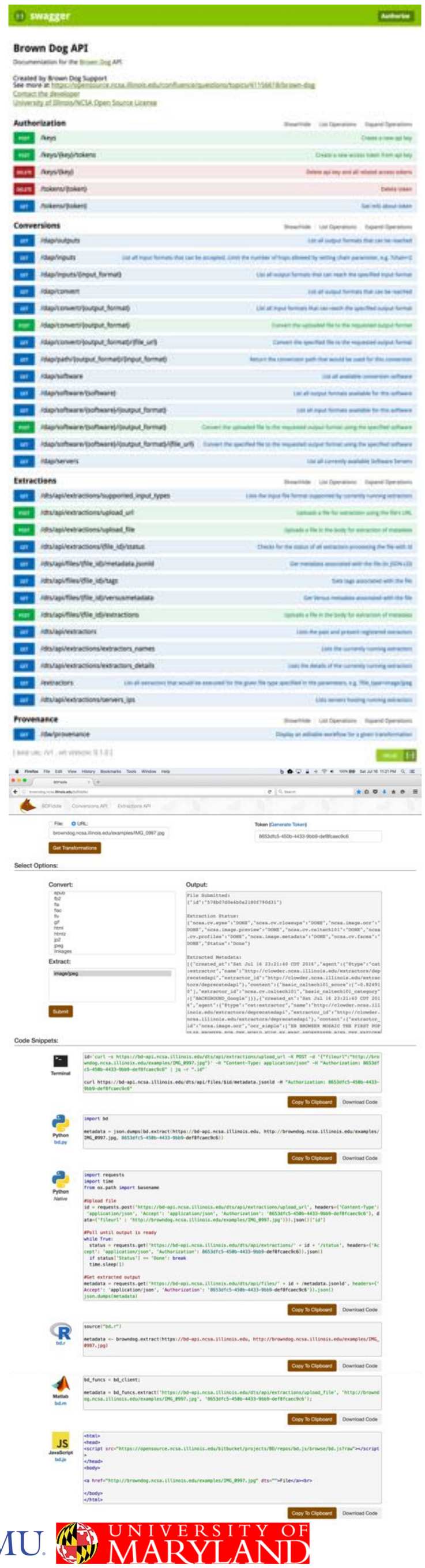
Brown Dog – A Science Driven Data Transformation Service

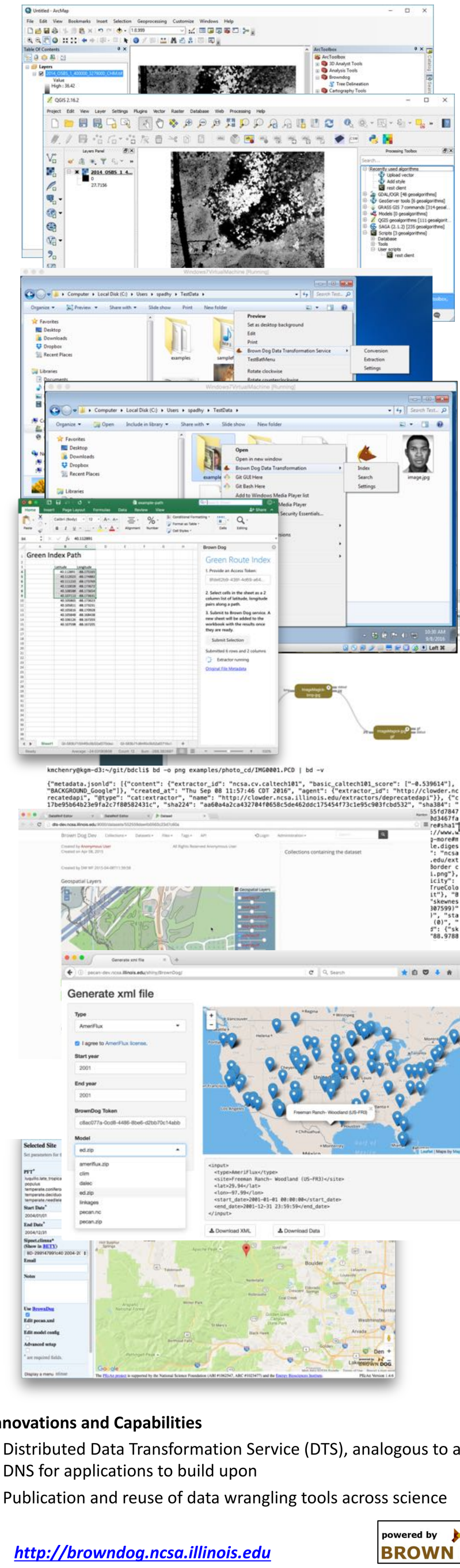
Kenton McHenry, Shannon Bradley, Mike Dietze, Praveen Kumar, Jong Lee, Richard Marciano, Luigi Marini, Jerome McDonough, Barbara Minsker, Art Schmidt, Bill Sullivan

Funded through National Science Foundation Cooperative Agreement ACI-1261582

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. Brown Dog aims to alleviate much 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 and content based extractions 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.

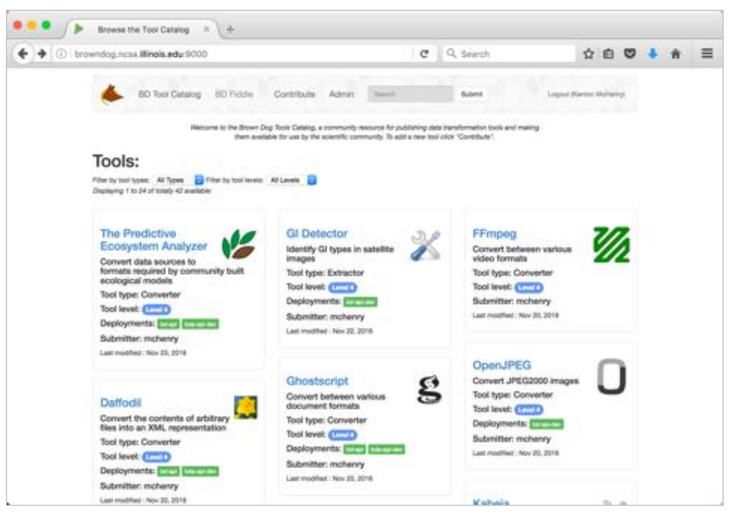






Domain	Tool	Type	Description
Biology	netcdf	Converter	Convert from binary netcdf to text
	PEcAn (Ameriflux)	Converter	Convert Ameriflux data to PEcAn's netcdf CF format
	PEcAn (DALEC)	Converter	Convert PEcAn's netcdf CF format to the format required by the DALEC model
	PEcAn (ED2)	Converter	Convert PEcAn's netcdf CF format to the format required by the ED model
	PEcAn (LINKAGES)	Converter	Convert PEcAn's netcdf CF format to the format required by
Hydrology	Advection Diffusion	Extractor	Solve a general advection-dispersion equation
	Chemical Mean Age	Extractor	Determine the mean age of chemical constituents with inputs of chemical dynamics
	Document Tables Extractor	Extractor	Extract tables from documents
	GDAL	Extractor	
	Historical River Extractor	Extractor	Extract the river networks from the ancient hand-drawing maps and compare them with current river networks
	Normalized Difference Vegetation Index	Extractor	
	River Chi Index	Extractor	Identify the river dynamics in a river basin and evaluate human activities' influences through Chi index in the streams
	River Sinuosity	Extractor	Study the maturity and equilibrium conditions of a stream
Green Infrastructure	Body of Water Detector	Extractor	Land coverage, extract locations of bodies of water from **satellite** data
	GI Identification	Extractor	
	Human Preference Score	Extractor	Assign a model derived human preference score to a given image of an urban environment
	Route Greenness	Extractor	Derive the green index of a city route
	Social Media GI	Extractor	
General	Tika	Extractor	Document extractions such as language identification, …
	txt2html	Converter	Convert text documents to HTML
	Versus – Color Distribution	Extractor	Generate a distribution of color values within an image to be used for comparing how similar two images are
	VLFeat	Extractor	Classify images as to whether they contain objects from the Caltech101 dataset (e.g. people, airplanes, motorcycles, cougars, …)
	Zip	Converter	Unzip zip archives
	Siegfried	Extractor	Extract information about a given file relevant to identifying its type and validating its format
	Stanford CoreNLP	Extractor	Natural Language Process extractions such as parts of speech, named entities, language, etc.
	Tesseract	Extractor	Object Character Recognition (OCR) to extract text from images containing text

82 Tools in beta release spanning Ecology, Biology, Hydrology, Civil Engineering, Social Science, as well as general users.



Innovations and Capabilities

- Distributed Data Transformation Service (DTS), analogous to a DNS for applications to build upon
- Publication and reuse of data wrangling tools across science

NCSA · ILLINOIS · UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · BOSTON UNIVERSITY · SMU · UNIVERSITY OF MARYLAND

http://browndog.ncsa.illinois.edu

powered by BROWN DOG