Research Data Challenges

• Storage
• Everything else!!!
  • The bytes are not enough on their own
    00110100 00110010
  • Metadata, curation tools, indexes, storage abstraction, replication, data transfer, authentication, access control, transformation, analysis, tools, computation, …
• Develop a deep symbiotic relationship between science and engineering users and developers of cyberinfrastructure to simultaneously advance new research practices and open transformative opportunities across all science and engineering fields.

• Provide an integrated and scalable cyberinfrastructure that leverages existing and new components across all areas of CIF21 and establishes a national data infrastructure and services capability.

• Ensure long-term sustainability for cyberinfrastructure, via community development, learning and workforce development in CDS&E and transformation of practice.

http://www.nsf.gov/cif21/
Architectural Vision for Research Cyberinfrastructure

Science Portals
- Community Metadata Improvement #1443062
- Local Spectroscopy Data Infra. #1640899
- Ocean Cloud Commons #1640775
- Whole Tale #1541450
- Modular Eng/Sci Cyber-platform #1443027

Applications & Frameworks
- Mobile Sensor Data #1640813
- Archiving U-Series Geochronologic Data #1443037
- Continuous Capture of Metadata #1640575
- Virtual Data Collaboratory #1640834
- Vizier #1640864
- Transient Data Access #1443083
- LearnSphere #1443068
- Collaborative Workflow Design #1443099
- Spatial Data Synthesis #1443080

Research Facilities
- Computer Aided Discovery in Geo #1442997
- Materials Engineering Data Lab #1640867
- ClearEarth #1443085
- Continuous Capture of Metadata #1640575
- HUBzero Geospatial #1261727
- Virtual Info. Fabric Infrastructure #1640818
- Pacific Research Platform #1541349
- Gravitational-Wave Workflow #1443047

Integrative Services
- DataOne
- PADS
- XSEDE
- Science Grid
- Agave

Resources
- NSF Resources
- Commercial Resources
- University Resources
- International Resources

STORM #1443046
- Brown Dog #1261582
- User Driven Architecture #1443070
- Data Exacell #1261721
- 4CeeD #1443013
- North East Storage Exchange #1640831
- Aristotle Cloud Federation #1541215
- Confidential Social Science Data #1443014

https://dibbs17.org
Smarr Taxonomy of Research CI Components

- **Data Applications**
  - Particles, Materials, Astro, Geo, GIS, Bio, Social, Environ, Ag, Medical, Sensors, etc

- **Data Cyberinfrastructure**
  - Computing, Storage, Federation, Clouds, Networking, SDN

- **Data Trust, Security, and Privacy**

- **Data Curation**
  - Capture, Annotation, Documentation, Archiving, Libraries, Management, Publishing

- **Data Discovery and Exploration**
  - Semantics, Ontology, Metadata, Data Mining, Web, Search

- **Data Sharing Middleware**
  - Accessibility, Collaboration, Hubs, Repositories

- **Data Workflows**

- **Data Analytics and Analysis**
  - Data-Intensive Computing, Machine Learning, NLP, Statistics
Brown Dog - A Science Driven Data Transformation Service

- Extensibility
  - Easy to add new transformations (i.e. converters and extractors)
  - Encapsulated transformation software & dependencies
- API
  - Supporting other applications/frameworks to build on top of
- Support for diverse usage (i.e. clients, languages, community tools & applications)
- Scalability, Distributed, Data Movement, Provenance with File Validation & Information Loss, Tool Preservation & Publication, Open Source

https://en.wikipedia.org/wiki/Mongrel
Brown Dog


curl -s https://bd-api.ncsa.illinois.edu/v1/extractions/url/ -X POST -d "{"fileurl":"http://browndog.ncsa.illinois.edu/examples/IMG_0997.jpg"}" -H "Content-Type: application/json" -H "Authorization: e6dab924-04c8-45c0-94aa-f0608c3c1a45" | jq -r ".id"
Selected Site
Set parameters for the run.

PFT
populus
temperate.coniferous
temperate.deciduous

Start Date
2004/01/01

End Date
2004/12/31

Signet.climna
Use AmeriFlux

Email

Use BrownDog

Edit pecan.xml

Edit model config

Advanced setup

Type
AmeriFlux

I agree to AmeriFlux license.

Start year
2001

End year
2001

BrownDog Token
c8ac077a-0d8-4486-8b86-d26b70c14abc

Model
ed.zip
ameriflux.zip
clim
disc
ed.zip
linkages
pecan.nc
pecan.zip

powered by
BROWN DOG
Geospatial Software
## Green Route Index

1. Provide an Access Token: 
   8f6762eb-436f-4a95-a64b...

2. Select cells in the sheet as a 2 column list of latitude, longitude pairs along a path.

3. Submit to Brown Dog service. A new sheet will be added to the workbook with the results once they are ready.

**Submit Selection**

Submitted 6 rows and 2 columns

**Adding metadata to sheet**

- Original File Metadata
- Derived File 0 metadata
- Derived File 1 metadata
- Derived File 2 metadata
- Derived File 3 metadata
- Derived File 4 metadata
- Derived File 5 metadata
- Derived File 6 metadata
- Derived File 7 metadata
- Derived File 8 metadata
- Derived File 9 metadata
- Derived File 10 metadata
- Derived File 11 metadata

---

<table>
<thead>
<tr>
<th>Route</th>
<th>Lat</th>
<th>Long</th>
<th>Green Index</th>
<th>Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40.11123</td>
<td>-88.17377</td>
<td>30.8</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>40.11123</td>
<td>-88.17377</td>
<td>37.7</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>40.11123</td>
<td>-88.17377</td>
<td>38.9</td>
<td>180</td>
</tr>
<tr>
<td>4</td>
<td>40.11123</td>
<td>-88.17377</td>
<td>27.6</td>
<td>270</td>
</tr>
<tr>
<td>5</td>
<td>40.11202</td>
<td>-88.17468</td>
<td>32.1</td>
<td>270</td>
</tr>
<tr>
<td>6</td>
<td>40.11202</td>
<td>-88.17468</td>
<td>10.9</td>
<td>180</td>
</tr>
<tr>
<td>7</td>
<td>40.11202</td>
<td>-88.17468</td>
<td>47.1</td>
<td>90</td>
</tr>
<tr>
<td>8</td>
<td>40.11266</td>
<td>-88.17517</td>
<td>31.6</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>40.11266</td>
<td>-88.17517</td>
<td>16.9</td>
<td>90</td>
</tr>
<tr>
<td>10</td>
<td>40.11266</td>
<td>-88.17517</td>
<td>19.9</td>
<td>270</td>
</tr>
<tr>
<td>11</td>
<td>40.11266</td>
<td>-88.17517</td>
<td>24.1</td>
<td>180</td>
</tr>
<tr>
<td>Lines of Code</td>
<td>273</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other files</td>
<td>Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Image Sample</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependencies</td>
<td>numpy, argparse, glob, cv2, cPickle, random, h5py, skimage, sklearn, scipy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficulties</td>
<td>Install OpenCV (cv2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total Code from 2 Files
<table>
<thead>
<tr>
<th>Lines of Code</th>
<th>47</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other files</td>
<td>None</td>
</tr>
<tr>
<td>Dependencies</td>
<td>bd, requests, os, glob, argparse, time, json, PIL</td>
</tr>
<tr>
<td>Difficulties</td>
<td></td>
</tr>
</tbody>
</table>

Total Code from 1 File
The Illinois Data Bank is a public access repository for publishing research data from the University of Illinois at Urbana-Champaign.

You are ready to deposit data if:

- your data is in a final state and not expected to undergo revisions.
- you have removed any private, confidential, or other legally protected information from your data.
- you are a faculty member, staff member, or graduate student at the University of Illinois at Urbana-Champaign.
- you have permission to publicly distribute data from all creator(s) and/or copyright owner(s).

Learn how to publish your data

Published data:

- is open to anyone in the world.
- receives a stable identifier (DOI) for easy reference and citation.
- is readily available for anyone to access for a minimum of 5 years.
- is located in a stable environment that complies with many funder and publisher requirements.

Review our policies

The Illinois Data Bank is a product of the Research Data Service at the University Library. See our Access and Use Policy. Contact us for questions and to provide feedback.
<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>160419TASSEL_results.txt</td>
<td>2.27 MB</td>
</tr>
<tr>
<td>2012_Collection_of_Poaceae_Germplasm_in_Eastern_Russia_Book_2.pdf</td>
<td>8.66 MB</td>
</tr>
<tr>
<td>2012_Collection_of_the_Poaceae_Germplasm_in_Eastern_Russia.pdf</td>
<td>12.4 MB</td>
</tr>
<tr>
<td>Dataset_S1-2012_COLLECTION_OF_MISCANTHUS_GERMLPLASM_IN_EASTERN_RUSSIA.csv</td>
<td>106 KB</td>
</tr>
<tr>
<td>GWASrussia160426_rrBLUP.R</td>
<td>14.1 KB</td>
</tr>
<tr>
<td>HapMapMsa.fas.txt</td>
<td>4.71 KB</td>
</tr>
<tr>
<td>HapMapMsa.hmc.txt</td>
<td>37.5 MB</td>
</tr>
<tr>
<td>HapMapMsa.hmp.txt</td>
<td>17.5 MB</td>
</tr>
<tr>
<td>HapMapMsi.fas.txt</td>
<td>4.21 MB</td>
</tr>
<tr>
<td>HapMapMsi.hmc.txt</td>
<td>31.5 MB</td>
</tr>
<tr>
<td>HapMapMsi.hmp.txt</td>
<td>70.5 MB</td>
</tr>
<tr>
<td>MiscanthusTagDatabase160224.xlsx</td>
<td>5.36 MB</td>
</tr>
<tr>
<td>Photo_locations_150810.csv</td>
<td>629 Bytes</td>
</tr>
</tbody>
</table>
Clowder (2013-Present)

NSF Innovative Systems and Software: Applications to NARA Research Problems (OCI-0525308)

Welcome to Clowder

Welcome to Clowder, a scalable data repository where you can share, organize and analyze data. This is a demo instance to try the system out. Please do not use this instance to store real data. We delete the content of this instance when we need to and it does not have very much disk space available. Thank you.