Posted on 25 Nov 2022 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.1002/essoar.10510197.1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Suite of Tools to Disseminate Information Derived from Multiple Domains Science Results

Drew Resnick[1], Xandre Chourio[2], Rémi Cousin[3], Kevin Graaf[2], Kyle Hall[2], Aaron Kaplan[2], Igor Khomyakov[4], and Audrey Vadillo[2]

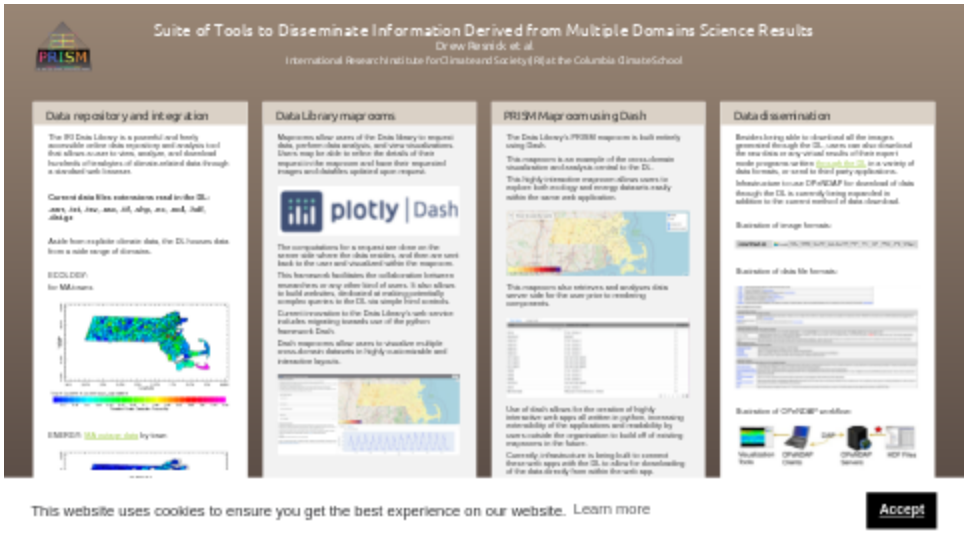
[1]Columbia University International Research Institute for Climate and Society
[2]International Research Institute for Climate and Society, Columbia University
[3]International Research Institute for Climate and Society (IRI), Columbia University
[4]Columbia University

November 25, 2022

Abstract

We present here a set of tools to curate, analyze, visualize and disseminate datasets, as well as communicate scientific findings, through easy-to-use web interfaces. PRISM uses the IRI Data Library (DL) to aggregate data from climate, agriculture, hydrology, ecology, finance, power outage and space weather domains of science, as well as the datasets resulting from their interdisciplinary research. These datasets vary in their spatial and temporal (and other) representations that the DL allows however to consistently represent in a multi-dimensional data framework. Proper curation facilitates interoperability and dissemination via other tools. We thus use popular python libraries (e.g. xarray, numpy, pandas) to query, manipulate and visualize the data and then build web applications to facilitate their exploration. The content and the interface of those applications, which we call Maprooms, are co-developed with the PRISM scientists in order to design the most relevant way to communicate their results.

Suite of Tools to Disseminate Information Derived from Multiple Domains Science Results



Drew Resnick et al

International Research Institute for Climate and Society (IRI) at the Columbia Climate School

PRESENTED AT:

DATA REPOSITORY AND INTEGRATION

The IRI Data Library is a powerful and freely accessible online data repository and analysis tool that allows a user to view, analyze, and download hundreds of terabytes of climate-related data through a standard web browser.

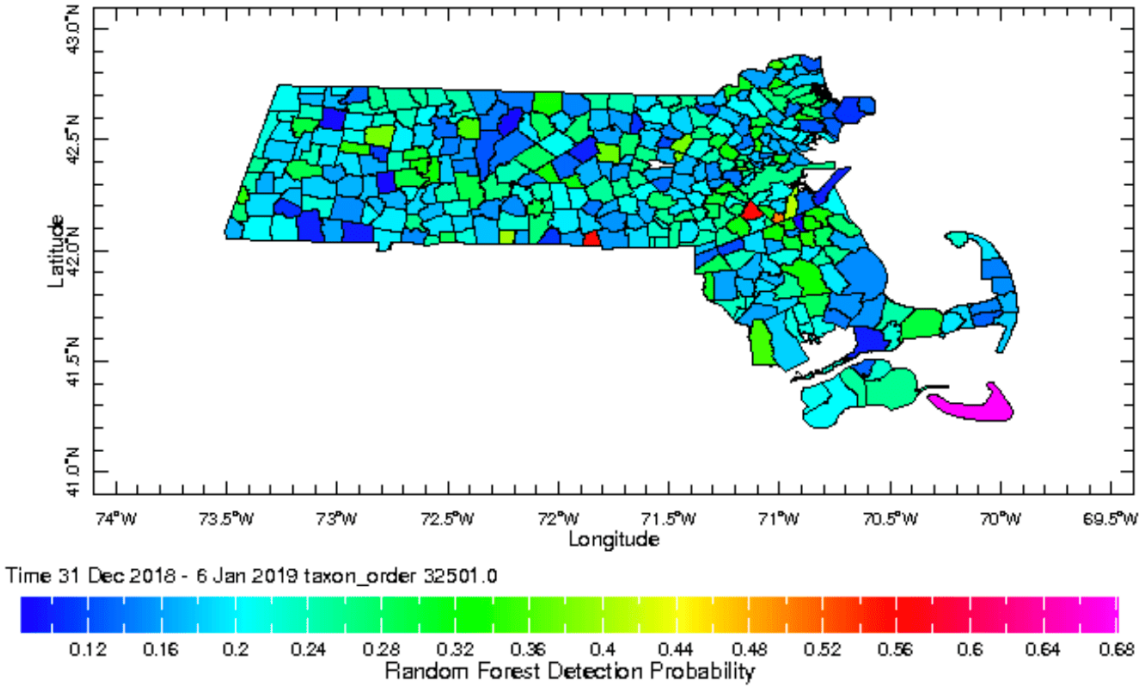
Current data files extensions read in the DL:

.zarr, .txt, .tsv, .asc, .tif, .shp, .nc, .nc4, .hdf, .dat.gz

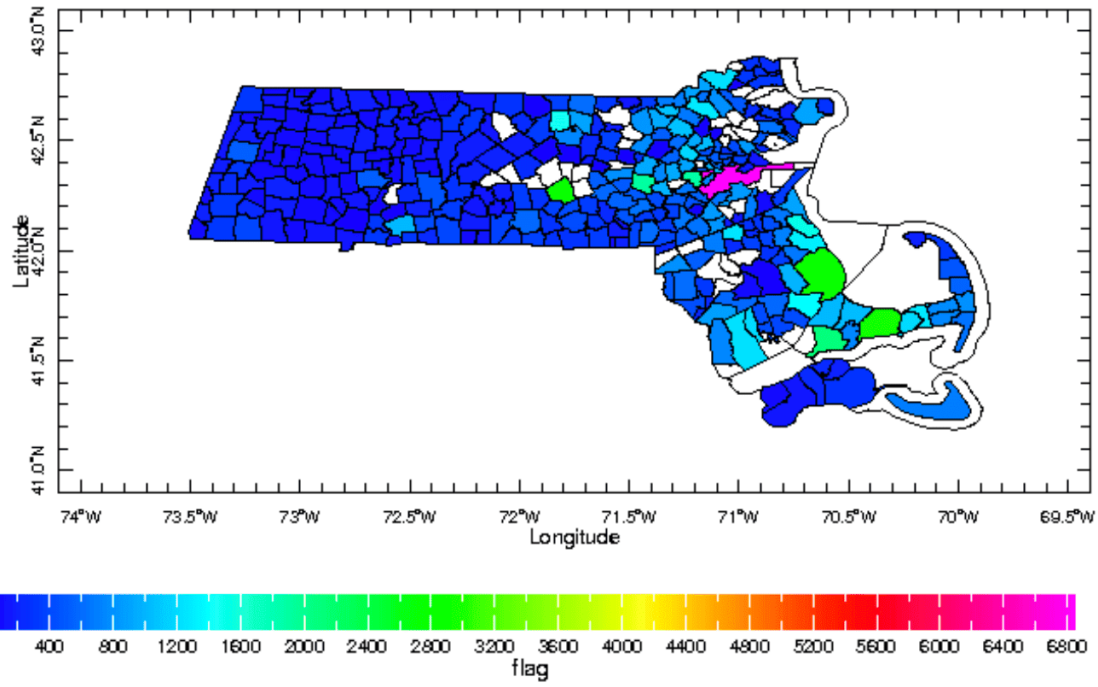
Aside from explicite climate data, the DL houses data from a wide range of domains.

ECOLOGY:

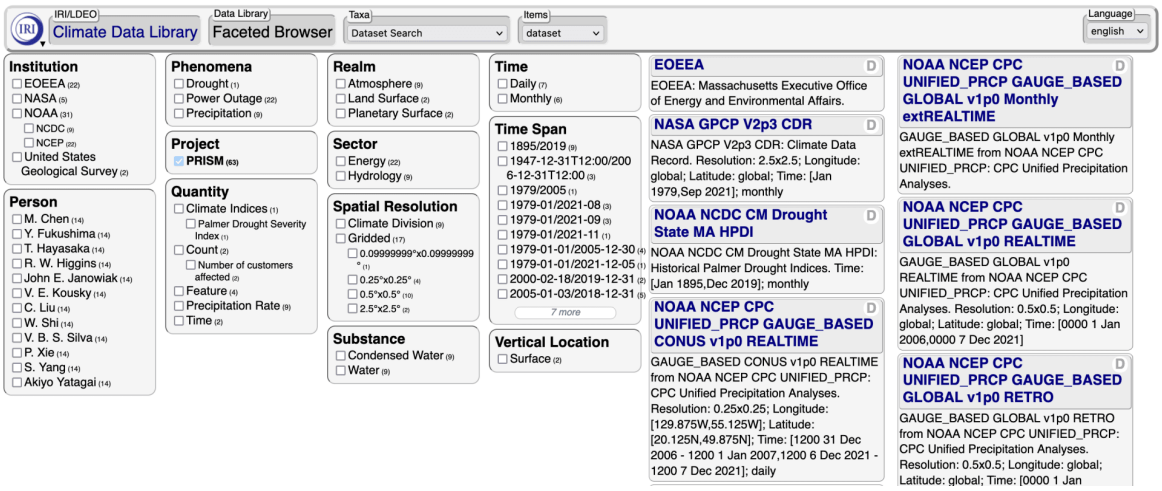
for MA towns



ENERGY: MA outage data (http://iridl.ldeo.columbia.edu/SOURCES/.EOEEA/a-/.SOURCES/.WORLDBATH/.bath /X/-74/-69.5/RANGE/Y/41/43/RANGE/-1/min/-a-/.the_geom/-a-/.ncacm/0/flagge/%5Bevent_id%5Dsum/-a-/.the_geom /-a/X/Y/fig-/white/mask/fillby/black/stroke/-fig/#expert) by town



Faceted search illustrates the wealth of data domains covered (note that tagging of data is ongoing and is thus incomplete).



DATA LIBRARY MAPROOMS

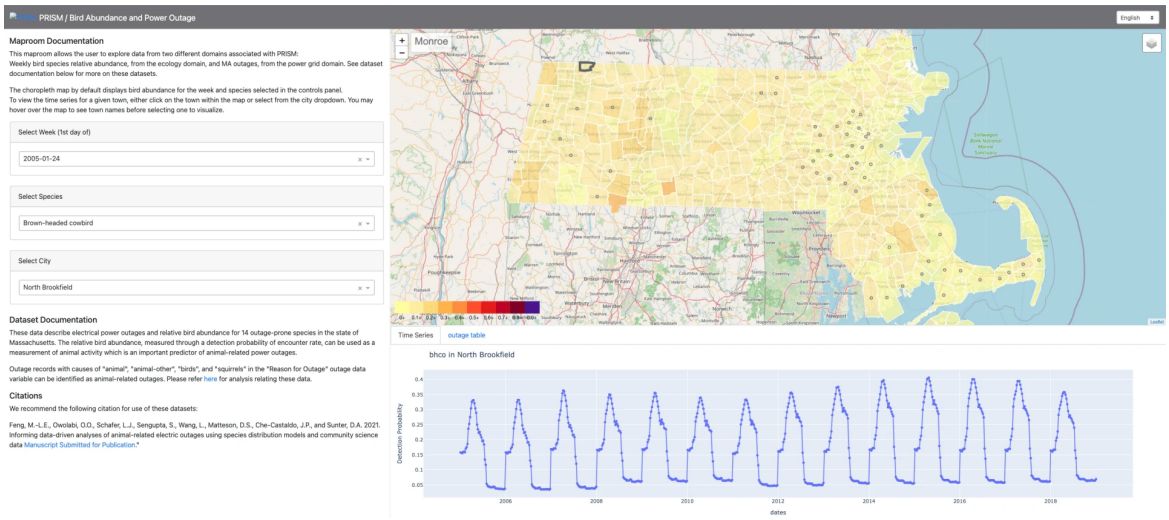
Maprooms allow users of the Data library to request data, perform data analysis, and view visualizations. Users may be able to refine the details of their request in the maproom and have their requested images and datafiles updated upon request.

The computations for a request are done on the server side where the data resides, and then are sent back to the user and visualized within the maproom.

This framework facilitates the collaboration between researchers or any other kind of users. It also allows to build websites, dedicated at making potentially complex queries to the DL via simple html controls.

Current innovation to the Data Library's web service includes migrating towards use of the python framework Dash.

Dash maprooms allow users to visualize multiple cross-domain datasets in highly customizable and interactive layouts.

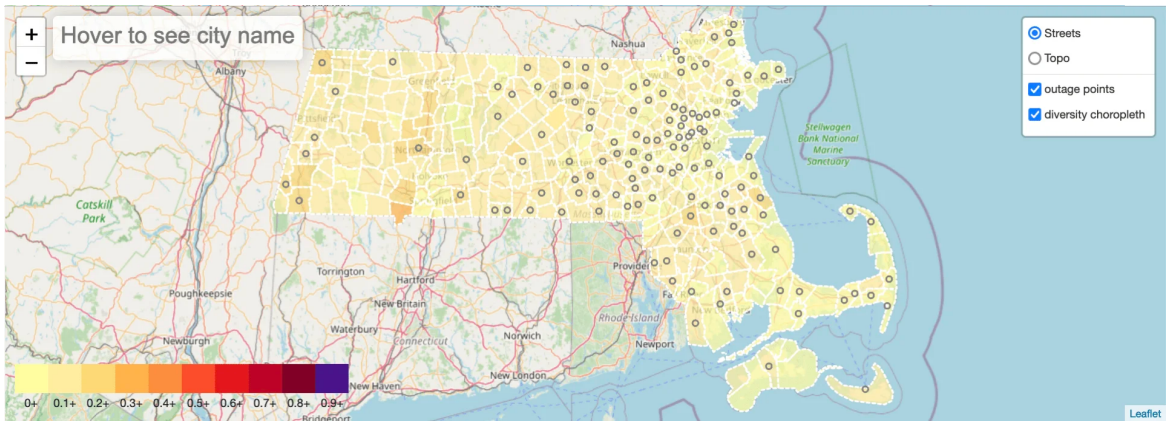


PRISM MAPROOM USING DASH

The Data Library's PRISM maproom is built entirely using Dash.

This maproom is an example of the cross-domain visualization and analysis central to the DL.

This highly interactive maproom allows users to explore both ecology and energy datasets easily within the same web application.



This maproom also retrieves and analyses data server side for the user prior to rendering components.

Time Series | outage table

city	reason_for_outage	outageCour
filter data...		
Acton	Tree Contact	3
Acushnet	Animal	1
Andover	Tree Contact	4
Andover	Tree Contact	4
Andover	Tree Contact	4
Andover	Tree Contact	4
Arlington	Failed Equipment	3
Arlington	Failed Equipment	3
Arlington	Failed Equipment	3
Ashby	Tree Contact	3
Ashby	Tree Contact	3
Ashby	Tree Contact	3
Ashland	Failed Equipment	1
Athol	Tree Contact	1
Belchertown	Physical Interference / Other	3

« ‹ 1 / 2 › »

Use of dash allows for the creation of highly interactive web apps all written in python, increasing extensibility of the applications and readability by users outside the organisation to build off of existing maprooms in the future.

Currently, infrastructure is being built to connect these web apps with the DL to allow for downloading of the data directly from within the web app.

DATA DISSEMINATION

Besides being able to download all the images generated through the DL, users can also download the raw data or any virtual results of their expert mode programs written through the DL (http://iridl.ldeo.columbia.edu/expert/) in a variety of data formats, or send to third party applications.

Infrastructure to use OPeNDAP for download of data through the DL is currently being expanded in addition to the current method of data download.

Illustration of image formats:

Illustration of data file formats:

ingrid	The Postscript-based software on which the Data Library is built.
CPT	Climate Predictability Tool More information
ferret	Interactive computer visualization and analysis software. More information
GrADS	Grid Analysis and Display System More information
matlab	Data analysis and visualization software. More information
NCL	NCAR Command Language More information
WinDisp	A public domain software package for the display and analysis of satellite images, maps and associated databases, with an emphasis on early warning for food security. More information

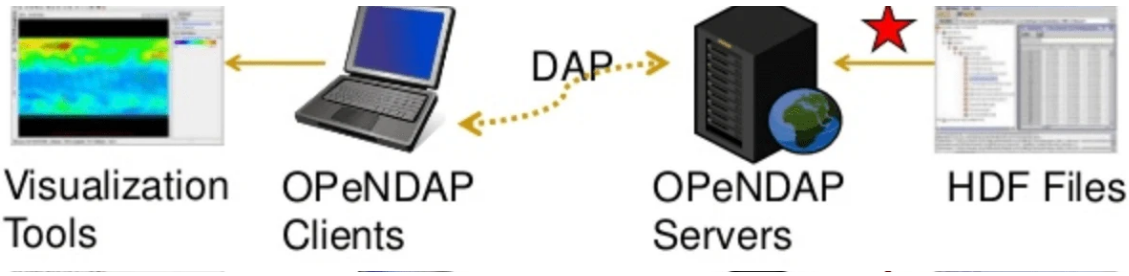
Other Available File Formats

Full Information Formats These files contain all of the available metadata.	
OPeNDAP	A system which downloads data directly to software, such as matlab, Ferret, GrADS, etc. Specific instructions are available in the table above. Note: OPeNDAP was formerly known as DODS (Distributed Oceanographic Data System). More Information
netCDF (network Common Data Form)	A commonly supported self-describing data format. More Information

Partial Information Formats These files contain only some of the available metadata.	
For the remaining data formats, the following information may be helpfulthe **scale_factor** is 1. and the **add_offset** is 0, i.e. the data is already properly scaled. The **missing_value** (flag for missing data) is NaN	
Columnar Table	A table with separate columns of numbers for each independent variable (i.e., grids) and for the data. This is an inefficient format, so you would have gotten a **HUGE** file for dataset of this size. This file will be approximately -1090676848 bytes, with 3 columns of 267024204 numbers.
2-Dimensional Tab-Separated Tables	Tab-separated-values (tsv) file with information about the independent variables (i.e., grids). The list to the left allows you to specify the format of the table. Note: The variable running across the top of the table (identifing columns) is listed first and the variable running down the side of the table (identifing rows) is listed second.
GIS-Compatible Formats There are three GIS-compatible formats available.	
2-Dimensional Table	A 2-dimensional ascii file that includes an ArcInfo Header.
IDA Image	File(s) in the Image Display and Analysis format. Typically used with WinDisp.
LAN Image	File(s) in the ERDAS LAN format. Typically used with various GIS programs, including ArcView and HealthMapper.
GeoTIFF Image	File in GeoTIFF format. Typically used with various GIS programs, including ArcView and ENVI.

Data Only Formats These files contain just the data without any of the available metadata.	
Binary direct access	A big-endian, ieee single-precision file in floating-point format. Also known as a binary random access file. This is a random-access file; it is purely data with no record-structuring information. The data is structured to correspond to the independent variables (i.e., grids) in X Y order, with the first grid varying the fastest.
DEC ALPHA direct access	Same as the binary random/direct access format above except that it is byte-swapped for DEC ALPHA's and PC's (little-endian).
Binary FORTRAN sequential access	A big-endian, ieee, single-precision file in floating-point format. This is a sequential-access file with each record containing all the X Y points. It **must** be read using FORTRAN sequential access. There is only one record for the data you have selected.
DEC ALPHA sequential access	Same as the binary sequential access format above except that it is byte-swapped for DEC ALPHA's and PC's (little-endian).
Text with tab-separated-values	Text file where data values corresponding to different **X** are separated by tabs and data values corresponding to different **Y** are on different lines. This is readable by most programs, including spreadsheets, but will be about four times larger than the binary or netCDF/HDF files noted above.
Text	Text file where data is arranged in chunks of X Y. There are five values per line and each chunk starts on a new line. This will be about four times larger than the binary or netCDF/HDF files.

Illustration of OPeNDAP workflow:



LINK TO A SURVEY

Enter your survey URL here

https://example.com

Submit

Delete