Requirements for a Biology node on ESGF

Patrik D'haeseleer
Sasha Ames
LLNL

ESGF can do much more than Climate Research

We've built this large federated database platform; what else can we use it for?

Many other research communities need a platform that can distribute Peta/Exa-scale data

- Physics
- Astronomy
- "Big Data" (e.g. social media interactions)
- **Biology**

Big Data in Biology

DNA sequencing technology is improving much faster than Moore's law! ~4X per year

~2000 sequencers, producing ~15 PB/yr (2013)

Illumina HiSeq X-10. Cluster of 10 sequencers; 2PB/yr

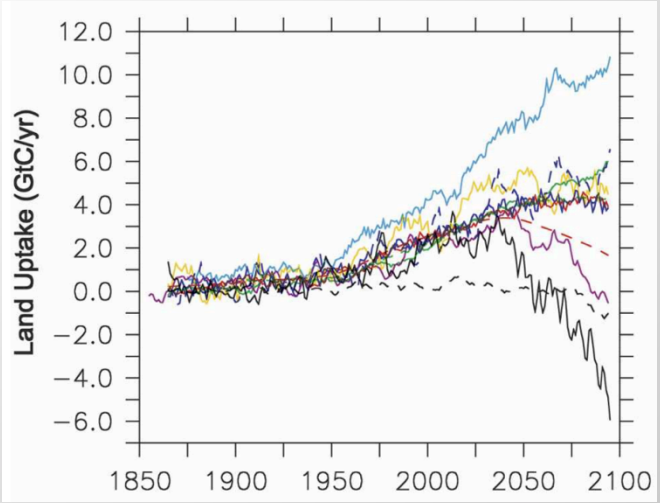
Plus **hundreds of other data types and formats**: genome annotations, metabolic models, protein structures, molecular dynamics simulations, medical records, ...

Any laboratory can be a data producer! "Do you really want 500 ESGF nodes?"

Synergies between Biology, Climate Research

1. Biogenic climate feedbacks

CMIP4 model predictions of biogenic soil C uptake/emission differ by more than the anthropogenic emissions!



2. **Epidemiology**: linking climate & disease

- Short-term predictions of future predictions based on past weather patterns
- Longer-term: predict shifting of tropical diseases towards temperate zones due to climate change

Large-scale spatiotemporal data that can be correlated with climate data on ESGF!

Sample epidemiology data sets

Project Tycho

- Disease outbreaks across the US, by week, since 1888
- Tier 1: 8 diseases, 1916-2009, 759,483 counts
- Tier 2: 47 diseases, 1888-2013, 3,418,529 counts

DengueDB Epidemiological Datasets

- Dengue outbreaks across the world, per year, since 1955

Dengue Virus Portal

- 2382 Dengue viral genomes, with information on year and country of isolation.

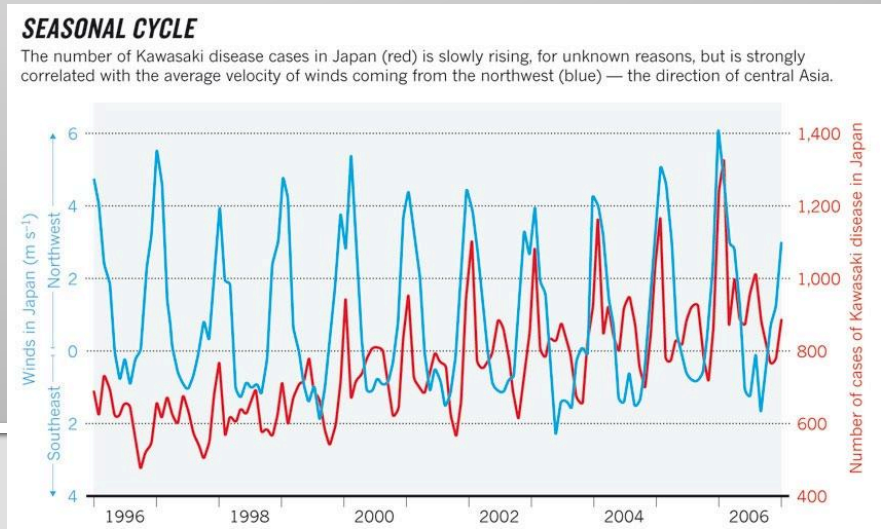
Applications

- Dengue is a tropical disease spread by mosquitos, now making inroads into US. Can we predict the future spread of Dengue due to global warming?
- Many diseases are correlated with weather patterns. Can we predict tomorrow's disease outbreaks based on last week's (or last year's) weather?

Climate variability and outbreaks of infectious diseases in Europe

Serge Morand[1,2], Katharine A. Owers[1], Agnes Waret-Szkuta[2], K. Marie McIntyre[3] & Matthew Baylis[3]

- Mystery "Kawasaki disease" in Japan may be due to fungal spores spread by high-altitude winds from central Asia



Issues with getting biological data into ESGF

hundreds of data types and formats; but almost nobody uses NetCDF

- Metadata may not be included in the file
- Fuzzy distinction between data & metadata
- Most data is not spatiotemporal
- Need different handlers for each data type
- Each data supplier may be using their own data formats

ESGF is surprisingly application agnostic, except project - model - experiment framework

Metadata convention for biological data

based on obs4CMIP5 Metadata Conventions

/<activity>/<product>/<data_type>/<agency>/
<project>/<organism>/ … /<version>/<filename>

- <activity> = "bio" (fixed)
- <product> = "observations" (fixed)
- <datatype> = “epidemiology” or “sequence”
- <agency> = the funding agency
- <project> = the funded project
- <organism> = bacteria or virus causing disease
- <version> = the dataset version, vYYYYMMDD

Epidemiological datasets:

/bio/observations/**epidemiology**/<agency>/<project>/
<organism>/**<variable>/<loc_type>/<frequency>**...

- <variable> = "cases", "deaths", "incidence", ...
- <loc_type> = “city”, “state”, or “country”
- <frequency> = "week", "year", ...

Sequence datasets:

/bio/observations/**sequence**/<agency>/<project>/
<organism>/**<seq_type>/<format>**...

- <seq_type> = “nucleotide”, “protein”
- <format> = “FASTA”, “FASTQ”, “genbank”,...


ESGF
Earth System Grid Federation
PCMDI
Home
Search
Tools
Account
Logout
Help
Current Selections
(x) activity:bio
Search Categories
Product
Activity
Type
epidemiology (8)
sequence (1)
Project
denguevirport (1)
tycho (8)
Agency
Broad (1)
NIH (8)
Organism
DIPHTHERIA (1)
HEPATITIS (1)
MEASLES (1)
MUMPS (1)
PERTUSSIS (1)
Search
Examples: temperature, "surface temperature", climate AND project:CMIP5 AND variable:hus.
To download data: add datasets to your Data Cart, then click on Expand or wget.
Temporal Search
Clear search constraints and datacart
Search Help
Search Controlled Vocabulary
Search All Sites
Show All Replicas
Show All Versions
< 1 > displaying 1 to 9 of 9 search results
Display 10 datasets per page
Add All Displayed to Datacart
Remove All Displayed from Datacart
Results
Data Cart
bio.tycho.DIPHTHERIA.tier1.city
Data Node: pcmdi11.llnl.gov
Version: 20141125
No description available.
Further options: Remove From Cart
bio.tycho.HEPATITIS.tier1.state
Data Node: pcmdi11.llnl.gov
Version: 20141125
No description available.
Further options: Add To Cart

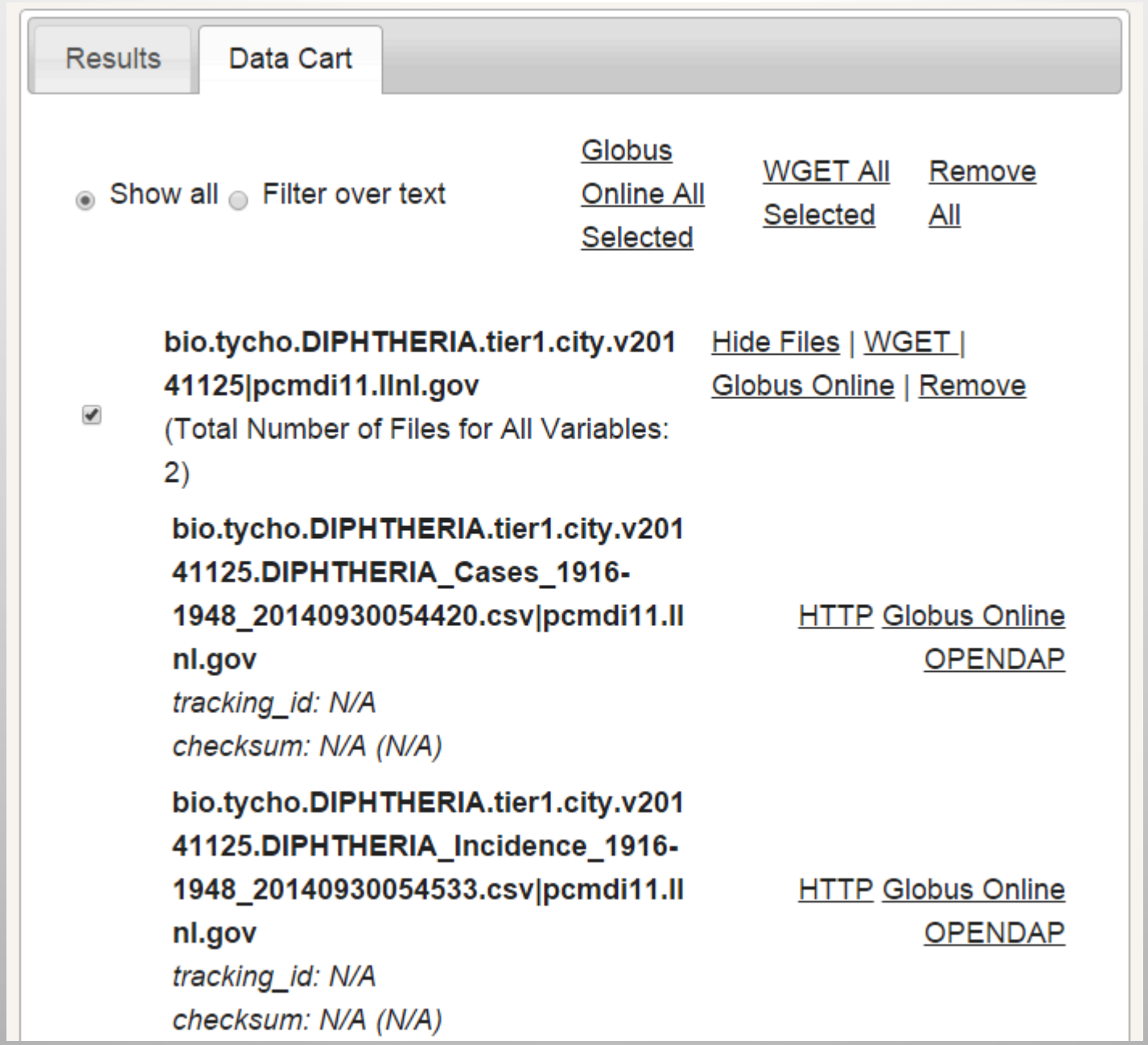
Dataset:

bio.tycho.DIPHTHERIA.tier1.city

Metadata

Show/Hide Properties access

	Property	Value
	access	GridFTP ; HTTPServer ; OPENDAP
	activity	bio
	agency	NIH
	cf_standard_name	cases ; incidence
	data_node	pcmdi11.llnl.gov
	data_type	epidemiology
collapse	dataset_id_template_	bio.%(project)s.%(organism)s.%(subset)s.%(loc_type)s
collapse	id	bio.tycho.DIPHTHERIA.tier1.city.v20141125|pcmdi11.llnl.gov
	index_node	pcmdi11.llnl.gov
	instance_id	bio.tycho.DIPHTHERIA.tier1.city.v20141125
	latest	true
	loc_type	city
	master_id	bio.tycho.DIPHTHERIA.tier1.city
	metadata_format	THREDDS
	number_of_aggregations	0
	number_of_files	2
	organism	DIPHTHERIA
	product	observations
	project	tycho
	replica	false
	score	1


Results
Data Cart
Show all
Filter over text
Globus Online All Selected
WGET All Selected
Remove All
bio.tycho.DIPHTHERIA.tier1.city.v20141125|pcmdi11.llnl.gov
(Total Number of Files for All Variables: 2)
Hide Files | WGET | Globus Online | Remove
bio.tycho.DIPHTHERIA.tier1.city.v20141125.DIPHTHERIA_Cases_1916-1948_20140930054420.csv|pcmdi11.llnl.gov
tracking_id: N/A
checksum: N/A (N/A)
HTTP Globus Online OPENDAP
bio.tycho.DIPHTHERIA.tier1.city.v20141125.DIPHTHERIA_Incidence_1916-1948_20140930054533.csv|pcmdi11.llnl.gov
tracking_id: N/A
checksum: N/A (N/A)
HTTP Globus Online OPENDAP

Yes, it's just a comma-separated table:

	A	B	C	D	E	F	G	H
2	Data provided by Project Tycho, Data Version 1.0.0, released 28 November 2013.							
3	YEAR	WEEK	BIRMINGHAM AL	MOBILE AL	MONTGOMERY AL	FORT SMITH AR	LITTLE ROCK AR	LOS ANGELES CA
4	1916	1	-	1	-	-	1	8
5	1916	2	2	-	-	-	-	4
6	1916	3	2	1	-	-	2	9
7	1916	4	-	-	-	-	1	12
8	1916	5	3	-	-	-	-	11
9	1916	6	-	1	-	-	-	17
10	1916	7	-	-	-	-	-	19
11	1916	8	-	-	-	-	-	12
12	1916	9	2	1	-	-	-	10
13	1916	10	1	-	-	-	-	19
14	1916	11	-	-	-	-	-	9
15	1916	12	-	-	-	-	1	10
16	1916	13	-	1	-	-	-	18
17	1916	14	1	-	-	-	-	9
18	1916	15	1	-	-	-	-	6
19	1916	16	-	-	-	-	-	6
20	1916	17	1	-	-	-	-	8
21	1916	18	-	-	-	-	-	3
22	1916	19	1	-	-	-	-	14
23	1916	20	1	-	-	-	-	13
24	1916	21	-	-	-	-	-	8
25	1916	22	1	-	-	-	-	13
26	1916	23	-	-	-	-	-	2

Could be turned into a NetCDF file, but that's not its native format

So, how can we actually DO something with this data?

This should not be just a data lookup service!

Server-side analytics is essential

including highly compute intensive search tools!

"find all the genes that are similar to this gene"

(like *"find all the weather patterns similar to this one"*)

For spatiotemporal datasets, can we convert them into NetCDF and upload into UV-CDAT? Should we need to?

Can/should we make data CF-compliant, where possible?

Can UV-CDAT import non-CF netCDF files? Should it?

CF Conventions mostly enforces self-documenting data, controlled vocabulary; little Climate & Forecast specific

Desiderata

Need GIS support for epidemiology data

- Cities, states, countries, regions
- Demographics (also for climate impact!)

Need much more powerful search than facets

- Search on numerical values and ranges
- Hierarchical categories (ontologies):
  - Organism taxonomy

    e.g “give me information on all viruses” → RNA viruses → ssRNA viruses → Filoviruses → Ebola strains
  - Open Biological & Biomedical Ontologies obofoundry.org

    e.g disease symptoms; environmental descriptors (terrestrial, aquatic, marine, benthic, sediment, …)

Work in progress!

Contact us if interested:

patrikd@llnl.gov