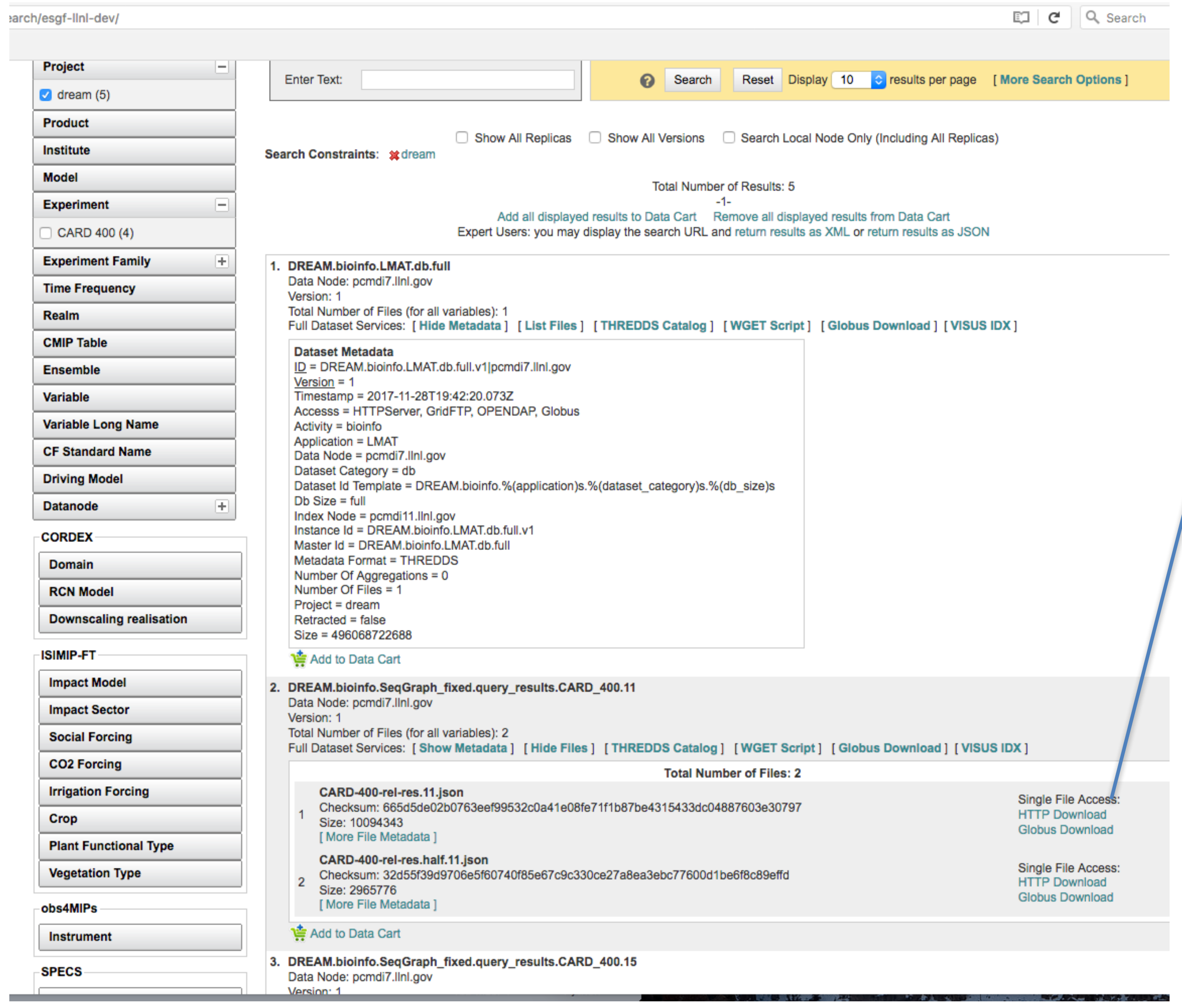
DREAM Data Services for Biological Data and Beyond

Sasha Ames, Luca Cinquini, and Dean Williams - DOE LLNL and NASA JPL

- One of the many goals of the DREAM project is to explore enhancements needed for ESGF to enable publishing and access to data in scientific fields other than climate/weather
- THREDDS data server (TDS) has been very effective for serving NetCDF data published to ESGF
- Need a service more specific for alternate data (e.g. ASCII-based) in other domains
    - Example: FASTA format used in bioinformatics to represent genomic and protein sequences
- This service will allow a variety of content types to interoperate properly with a user's web browser



earch/esgf-llnl-dev/

Project
☑ dream (5)
Product
Institute
Model
Experiment
☐ CARD 400 (4)
Experiment Family
Time Frequency
Realm
CMIP Table
Ensemble
Variable
Variable Long Name
CF Standard Name
Driving Model
Datanode

CORDEX
Domain
RCN Model
Downscaling realisation

ISIMIP-FT
Impact Model
Impact Sector
Social Forcing
CO2 Forcing
Irrigation Forcing
Crop
Plant Functional Type
Vegetation Type

obs4MIPs
Instrument

SPECS

Enter Text: Search Reset Display 10 results per page [More Search Options]

☐ Show All Replicas ☐ Show All Versions ☐ Search Local Node Only (Including All Replicas)

Search Constraints: dream

Total Number of Results: 5
-1-
Add all displayed results to Data Cart Remove all displayed results from Data Cart
Expert Users: you may display the search URL and return results as XML or return results as JSON

1. DREAM.bioinfo.LMAT.db.full
Data Node: pcmdi7.llnl.gov
Version: 1
Total Number of Files (for all variables): 1
Full Dataset Services: [Hide Metadata] [List Files] [THREDDS Catalog] [WGET Script] [Globus Download] [VISUS IDX]

Dataset Metadata
ID = DREAM.bioinfo.LMAT.db.full.v1|pcmdi7.llnl.gov
Version = 1
Timestamp = 2017-11-28T19:42:20.073Z
Accesss = HTTPServer, GridFTP, OPENDAP, Globus
Activity = bioinfo
Application = LMAT
Data Node = pcmdi7.llnl.gov
Dataset Category = db
Dataset Id Template = DREAM.bioinfo.%(application)s.%(dataset_category)s.%(db_size)s
Db Size = full
Index Node = pcmdi11.llnl.gov
Instance Id = DREAM.bioinfo.LMAT.db.full.v1
Master Id = DREAM.bioinfo.LMAT.db.full
Metadata Format = THREDDS
Number Of Aggregations = 0
Number Of Files = 1
Project = dream
Retracted = false
Size = 496068722688

Add to Data Cart

2. DREAM.bioinfo.SeqGraph_fixed.query_results.CARD_400.11
Data Node: pcmdi7.llnl.gov
Version: 1
Total Number of Files (for all variables): 2
Full Dataset Services: [Show Metadata] [Hide Files] [THREDDS Catalog] [WGET Script] [Globus Download] [VISUS IDX]

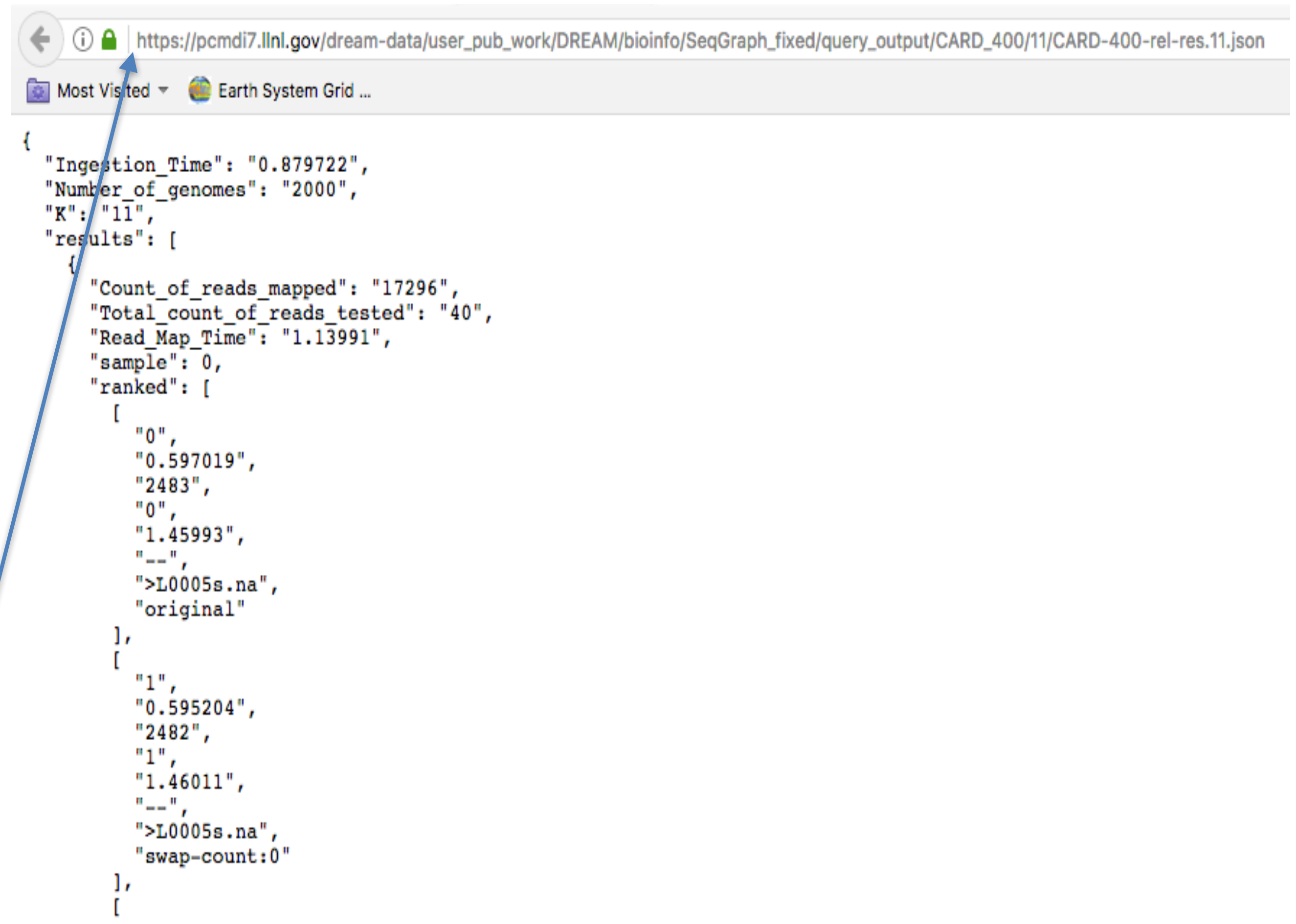
Total Number of Files: 2

1 CARD-400-rel-res.11.json
Checksum: 665d5de02b0763eef99532c0a41e08fe71f1b87be4315433dc04887603e30797
Size: 10094343
[More File Metadata]
Single File Access: HTTP Download Globus Download

2 CARD-400-rel-res.half.11.json
Checksum: 32d55f39d9706e5f60740f85e67c9c330ce27a8ea3ebc77600d1be6f8c89effd
Size: 2965776
[More File Metadata]
Single File Access: HTTP Download Globus Download

Add to Data Cart

3. DREAM.bioinfo.SeqGraph_fixed.query_results.CARD_400.15
Data Node: pcmdi7.llnl.gov
Version: 1



https://pcmdi7.llnl.gov/dream-data/user_pub_work/DREAM/bioinfo/SeqGraph_fixed/query_output/CARD_400/11/CARD-400-rel-res.11.json

Most Visited Earth System Grid ...

```
{
  "Ingestion_Time": "0.879722",
  "Number_of_genomes": "2000",
  "K": "11",
  "results": [
    {
      "Count_of_reads_mapped": "17296",
      "Total_count_of_reads_tested": "40",
      "Read_Map_Time": "1.13991",
      "sample": 0,
      "ranked": [
        [
          "0",
          "0.597019",
          "2483",
          "0",
          "1.45993",
          "--",
          ">L0005s.na",
          "original"
        ],
        [
          "1",
          "0.595204",
          "2482",
          "1",
          "1.46011",
          "--",
          ">L0005s.na",
          "swap-count:0"
        ],
        [
```

New tools:

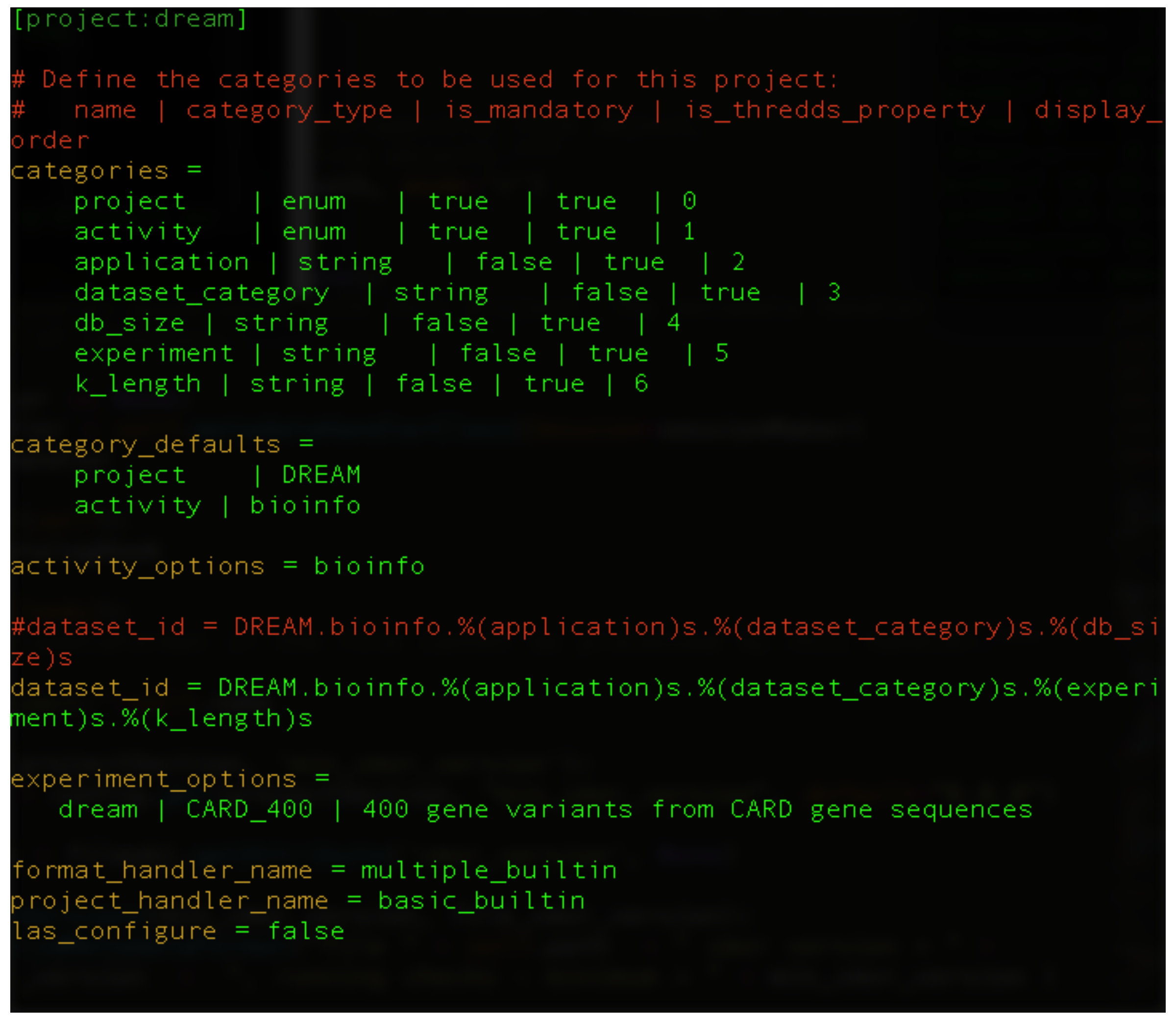
- get_meta.sh
- create_dset.py

  - scripts to generate a generic mapfile with consistent directory_format to dataset_id mapping
- dream-data
    - Service supplies correct mime times for published files
    - eg. images, .pdf, .json
    - Flask module
    - https://github.com/ESGF/esgf-dream-data-service

Steps taken for publishing to service:

- Modify esg.ini – use /dream-data/ for HTTPServer
- esg.dream.ini – use multiple_handler to accept non-netcdf data
- Generate mapfile with get_meta.sh + create_dset.py
- Reconfigure DRS when sub-project config changes, repeat

esg.dream.ini`



```
[project:dream]

# Define the categories to be used for this project:
#   name | category_type | is_mandatory | is_thredds_property | display_
order
categories =
    project     | enum    | true  | true  | 0
    activity    | enum    | true  | true  | 1
    application | string    | false | true   | 2
    dataset_category  | string    | false | true   | 3
    db_size | string    | false | true   | 4
    experiment | string    | false | true   | 5
    k_length | string | false | true | 6

category_defaults =
    project     | DREAM
    activity | bioinfo

activity_options = bioinfo

#dataset_id = DREAM.bioinfo.%(application)s.%(dataset_category)s.%(db_si
ze)s
dataset_id = DREAM.bioinfo.%(application)s.%(dataset_category)s.%(experi
ment)s.%(k_length)s

experiment_options =
   dream | CARD_400 | 400 gene variants from CARD gene sequences

format_handler_name = multiple_builtin
project_handler_name = basic_builtin
las_configure = false
```

Benefit to input4MIPs: Example MPI-M "dataset" has .pdf, README, and code files

Future work:

- Add random access (FASTA and JSON)
- Content listing in service
- Support formats with metadata to be extracted (need to determine which)
- JSON input integrated into publisher for attribute and map specification
- Manage config externally (retire ini) using CV services and support mutable DRS projects

LLNL-POST-742505

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC