Developing an integrated agriculture data management system

Engaged research teams often times generate large amounts of diverse, heterogeneous data. Yet the integration and use of such data in a collaborative manner can be difficult. The areas of data management and how best to envision integrated, yet heterogeneous, research collaboration have spawned several models for data storage and analysis (Chernenk 2006, Papazoglou, 2004). REACH PNA – a USDA funded coordinated agriculture project to explore climate impacts on crop production systems in the Pacific northwest, has developed a data management system for these areas of heterogeneous science data integration.

REACH PNA DM System: Integrated Data Science Architecture

The REACH data management effort has a focus to develop modular, sustainable, and extensible systems/processes that would allow for the collection, storing, and analyzing of REACH-related data and content in a transdisciplinary manner. In support of this strategy, we have built five core integrated systems to implement this approach:

1. Our [http://www.reacchpna.org](http://www.reacchpna.org) portal;
2. the REACH Data Library for data uploading/searching (http://data.reacchpna.org);
3. the REACH Analysis Library for analytical tools (http://analysis.reacchpna.org);
4. a THREDDS Data Catalog for array based, meteorological and climate scenario data (http://thredds.reacchpna.org); and

Integrating data using RESTful web services

As mentioned above, the data management methodology is based on the development of systems that are modular, sustainable after the life of the project, and extensible – with regards to interacting with other systems and processes, as well as usable by researchers and the public at large. In addition, we use Representational State Transfer Protocol (REST) to display and provide query functionality to data stored in our REACH Data Library, as well as our THREDDS server, which contains array-based historical and future climate scenario datasets. As an analytic ingestion tool – IPython (Interactive Python) serves as a mechanism to access and combine datasets.

Interactive Python Notebook Server. Interactive Python, or Jupyter, is a somewhat new development over the last two years, to enable the compilation of Python within a web browser. This shared notebook model provides a new way for scientific researchers to collaborate, in real-time, on data analysis and interrogation using Python ([https://jupyter.reacchpna.org](https://jupyter.reacchpna.org)).

THREDDS. Thematic Realtime Environmental Data Distribution Services (THREDDS), is a data cataloging approach that has been developed UNIDATA, a group that is part of the University Consortium of Atmospheric Research (UCAR). THREDDS is a Java-based server technology that is used for dissemination, aggregation, and sub-setting of multi-variable data, such as NetCDF formatted datasets.

REACH Data Policy Agreements

The REACH Data Policy and Data Agreement is an important aspect of the overall project. Defining and describing the policies that regulate data contribution, data management, as well as the protocols and procedures that researchers will abide by regarding data collaboration – is extremely important.

Leveraging the solution at a national level: National Ag Data Network Harmonization

The REACH team is engaged with a number of collaborators at the University of Florida, and other institutions, to establish a USDA based national data network for agriculture.

REACH Data Library

The REACH Data Library, accessed from the REACH web portal, is the core location to access REACH data. Implemented using ESRI’s Geoportal Server, running on Linux and using a PostgreSQL geospatial database, the Data Library is a common location for heterogeneous data access.

REACHPNA.ORG Web Portal

The REACH Web Portal ([https://www.reacchpna.org](https://www.reacchpna.org)) is the central entry point for all public and secure information. REACH members access the portal using a secure login and password, which in turn provides varying access to data uploading, searching, and analysis tools.

REACH Data Policy Agreements

The REACH Data Policy and Data Agreement is an important aspect of the overall project. Defining and describing the policies that regulate data contribution, data management, as well as the protocols and procedures that researchers will abide by regarding data collaboration – is extremely important.

Leveraging the solution at a national level: National Ag Data Network Harmonization

The REACH team is engaged with a number of collaborators at the University of Florida, and other institutions, to establish a USDA based national data network for agriculture.