“Assessing the value and use of data repositories for integrated research efforts”

Pacific Northwest Climate Science Conference – September 9th & 10th, 2014 - Seattle, WA

Erich Seamon, M.S. PMP GISP, - REACCH Environmental Data Manager, College of Agricultural and Life Sciences, University of Idaho erichs@uidaho.edu

Paul Gessler, Ph.D. - Professor, Department of Forest, Rangeland, and Fire Sciences, College of Natural Resources, University of Idaho

Von Walden, Ph.D. – Professor, Department of Civil and Environmental Engineering, Washington State University

Edward Flathers, M.S. - Department of Forest, Rangeland, and Fire Sciences, College of Natural Resources, University of Idaho

Stephen Fricke, M.S. – REACCH Programmer, College of Agricultural and Life Sciences, University of Idaho

http://bitly.com/seamon2014poster

Data Management & Repository Development Methodologies

Engaged research teams often times generate large amounts of data, in varying forms, as part of their efforts. The areas of data management and how best to envision heterogeneous, geographically separated, research collaboration have spawned several models for data storage, curation, dissemination, and analysis (Chernenak 2000, Papazoglou, 2004). Three approaches that we examined include:

1) Federated System – using the DataONE model as an example. Systems and functionality are grouped and federated.
2) Distributed System. Data and systems are fully distributed and are not collated at one location. Application processes are put in place to dynamically organize and provide data as needed.
3) Independent Replication Systems. Processes and data are replicated and managed at separate grouped locations. Data may be synchronized on a periodic basis.

DataONE federated system example

Data Repository Case Study Overview - REACCHPNA

The REACCH Data Management effort has a focus to develop modular, sustainable, and extensible systems/processes that would allow for the collection, storing, and analyzing of REACCH-related data and content. In support of this strategy –we have built out four core systems to implement this approach:

1. Our http://www.reacchpna.org portal;
2. REACCH Data and Analysis Libraries;
3. A THREDDS Data Catalog; and
4. an Interactive Python Notebook Server.

Supporting these four core areas is a developed architecture that includes a three-tier server environment (data, applications, web), a metadata cataloging server (a customized version of ESRI’s Geoportal Server), a geospatial web server environment for web mapping services (ArcGIS Server), and a geospatial enterprise database (PostgresQL) – all interconnected to an LDAP server for unified user logins across systems. In addition, all data is replicated/mirrored @ Idaho National Laboratories (INL).

REACCH Data and Analysis Libraries

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.

Our REACCH metadata harvesting/cataloging approach is primarily based on the ISO 19115-2 metadata standard. With much of our agricultural-based data having a geographic structure, the ISO-19115-2 standard fits well with our needs over the long term. In addition, our methodology required the integration of other ‘soft’ datasets (publications, presentation, images) to be closely tied to related raw datasets. As such, we use the Dublin Core metadata standard for all non-geographic data, and use metadata lineage attributes to interconnect a publication or presentation with its’ related raw data. In this way, we provide a structure that links:

1. Quantifiable datasets, with
2. Related ‘soft’ data (publications, presentations, etc.) that are the result of raw data output (Figure 1).

In order to catalog datasets with varying metadata standards, we customized ESRI’s Geoportal Server software – which is open-source and java-based – running under Tomcat on Red Hat Linux. We incorporated an upload feature to allow users to not only add data, but to meta-tag those datasets in a secure fashion – thru LDAP user groups. We, in turn, transform uploaded datasets into web/map services (using ArcGIS Server) – and then register these web/map services back to the REACCH Data Library.

In addition, we use Representational State Transfer Protocol (REST) to display and provide query functionality to both data stored in our REACCH Data Library, as well as to query web/map services using the REST protocol. This approach is likewise usable with both our THREDDS and Ipython server approaches, as described below.

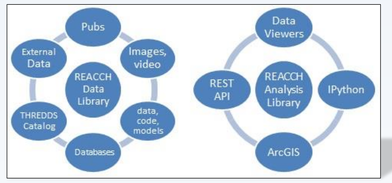

Figure 1. REACCH Data Interaction Diagram (2013)

REACCH THREDDS and Interactive Python Servers

In addition to the use of a Data Cataloging Library for metadata tagging and harvesting – REACCH has implemented/customized two additional server-based technologies for data interrogation and examination:

THREDDS. Thematic **R**ealtime **E**nvironmental **D**ata **D**istribution **S**ervices (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. REACCH has extensive climatic model runs and historical meteorological datasets that are in a NetCDF format – and in order to enable the interaction of these datasets with other data formats (point, polygon, or other geographic-based datasets) – we have implemented a THREDDS environment (https://thredds.reacchpna.org), with over 20 terabytes of climate and meteorological datasets. Several advantages of using THREDDS include:

1. The ability to use OPENDAP protocols for data interrogation using Python;
2. Ability to integrate THREDDS URLS with Interactive Python Notebooks for data examination and analysis; and
3. Easily subset very large NetCDF datasets based on geographic, time, or other meteorological or climate run variables.

Interactive Python Notebook Server. Interactive Python, or Ipython, 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://ipython.reacchpna.org). Ipython provides some excellent capabilities with regards to climatic science integration:

1. Ability to use REST protocols to examine data;
2. Use of Matplotlib to examine and analyze datasets in real-time using fairly simple python calls; and
3. Ability to interrogate NetCDF4 model data.

Interactive Python is developed by Fernando Perez of the University of Colorado at Boulder, and Brian E Granger of the Tech-X Corporation (http://fperez.org/papers/ipython07_pe-gr_cise.pdf).

Data Repository Case Study: REACCHPNA

REACCH Data Library

The REACCH Data Library is the core location for REACCH information storage. Implemented using ESRI’s Geoportal Server, running on Linux and using a PostgresQL geospatial database – the Data Library is accessed by our REACCH Analysis Library and data analysis tools.

Geoportal Server is a java-based web server technology, that runs under Tomcat – within a Linux Red Hat environment. Equipped with metadata editing capabilities – we have extended this software to enable file uploading, as well as using multiple customized metadata standards. REACCH currently uses the **ISO 19115-2** metadata standard for raw data, and the **Dublin Core** standard for publications and presentations.

Built upon the REST API protocol, Geoportal Server provides an excellent interface structure for accessing data and web services thru developed interfaces, or by simple URL constructs.

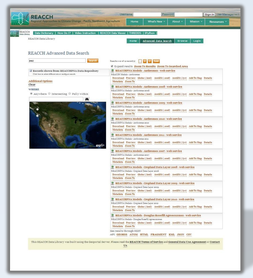

Figure 3. REACCH Data Library
https://www.reacchpna.org/resources/reacch-data-library

REACCH Analysis Library

The REACCH Analysis Library is a central location where we provide access to all of our technology-based analysis tools.

The REACCH Analysis Library is focused on providing analytical analysis capabilities using ArcGIS Server, Javascript, and python. Our analysis efforts can be described in three steps:

1. Data is uploaded to our data library, in varying formats;
2. That data is transformed, when possible, into geographic datasets and stored in our geospatially-enabled PostgresQL database;
3. These data layers are then exposed via a web/map service using ArcGIS server – that is then consumed and manipulated using javascript and python.

Our python development integrates with our mapping services thru the use of python-based geoprocessing services developed using the ArcGIS Server API. These toolsets are developed, and then ported to ArcGIS Server – where they interact within a javascript-based web browser to allow for data interrogation and examination of both PostgresQL and NetCDF datasets. In this way, we allow for the dynamic integration of geographic, database layers, with NetCDF climatic datasets. Examples of such interaction include:

1. Growing Degree Calculators;
2. Crop buffer tools; and
3. Biotic/Climatic queries.

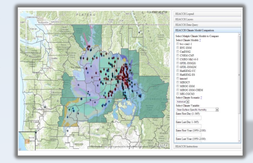

Figure 4. REACCH Analysis Library – the REACCH Analysis Library can be found under the Resources Tab. The REACCH Analysis Library has analysis viewers as a first effort in displaying REACCH information.

https://www.reacchpna.org/resources/reacch-analysis-library

REACCHPNA.ORG Web Portal

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

While seen as simply a ‘web site’, the reacchpna.org portal is a critical aspect of the REACCH project. With the use of technology being paramount – having a well managed and understandable web presence is very important. Some key components of the web site include:

1. Dynamic content added via a blog/rss style;
2. Facebook, YouTube, Twitter cross-over integration;
3. User login management that integrates with data access;
4. Cross-institutional accessibility;

We would encourage you to review our content at https://www.reacchpna.org.

Figure 2. https://reacchpna.org web portal

REACCHPNA System Architecture and Information Flow

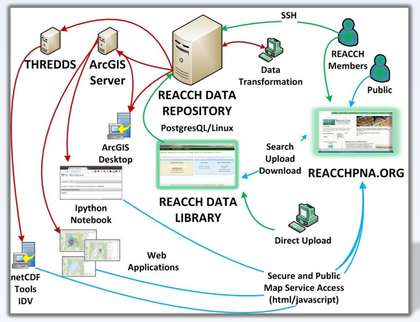

Figure 8. REACCH System Architecture

The REACCH data flow is outlined in Figure 8 above. As outlined in the previous section, the REACCH entry point for data information access is https://www.reacchpna.org. While there are other mechanisms to access our Data Library, informational entry using https://www.reacchpna.org is the more general approach.

Researchers upload data directly to the REACCH Data Library, where metadata is added and attached to the dataset for transformation and data insertion to the repository. Application and web servers interrogate the datasets from our Data Repository – and provide for analysis using;

- ArcGIS Desktop direct connection to the REACCH Data Repository
- Web Application Access using REACCH Web Mapping Services

Other functional datasets – including policy, documentation, and media – are collected and displayed as REACCH libraries – listed under the Resources Tab (after secure login).

REACCH Interactive Python Server

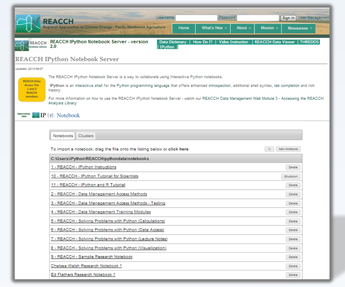

Figure 5. REACCH Interactive Python Server

The REACCH Interactive Python Server (Ipython – Figure 5) – is a server-side use of Interactive Python – exposed to REACCH members for collaboration and programming purposes. Interactive Python is developed by Fernando Perez at the University of Colorado at Boulder and Brian E. Granger from the Tech-X corporation (http://fperez.org/papers/ipython07_pe-gr_cise.pdf) – and is an excellent approach to enabling researchers to collaborate and interact with datasets using Python – in a web browser.

REACCH has set up an Ipython server with several notebooks – that can be accessed here:

https://ipython.reacchpna.org

The REACCH THREDDS Data Catalog (Figure 6) is a server-side software technology that aggregates very large datasets that cannot be stored in a database or the REACCH Data Library. THREDDS – or **Th**ematic **R**ealtime **D**ata **D**istribution **S**ervices – is a java-based technology that is developed by UNIDATA – a technology wing of the National Consortium of Atmospheric Research (UCAR).

REACCH stores over 20 terabytes of climatic-based datasets (NetCDF) with our THREDDS server, which allows for aggregation and sub-setting of data based on time, geography, and other climatic-based variables.

Our THREDDS server is exposed and consumed by our REACCH Data Library – so any THREDDS data can be accessed directly from our THREDDS interface, or by searching for this data from our REACCH Data Library.

You may directly connect to our THREDDS server here:

http://thredds.reacchpna.org

REACCH Data Policy and Data Agreement

The REACCH 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.

The REACCH Research Data Policy outlines several areas, including:

1. Our research policy statement on communication, accountability, timeliness, and expectations;
2. Federal and state compliance;
3. Describes the potential data types and usage;
4. Outlines our general data use agreement as well as our restricted data use agreement.

Each REACCH researcher and student is required to accept and comply with the overall REACCH Research Data Policy – which can be signed electronically thru our https://www.reacchpna.org portal.

The REACCH Research Data Policy – with addendum data agreements, can be seen here:

https://www.reacchpna.org/resources/reacch-policy-library

Figure 7. REACCH Research Data Policy

REACCH THREDDS Data Cataloging

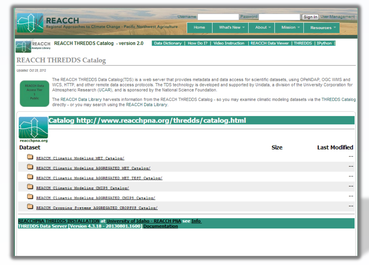

Figure 6. REACCH THREDDS Data Catalog Server

REACCH Data Management Partners

“Regional Approaches to Climate Change for Pacific Northwest Agriculture” is funded through award #2011-68002-30191 from the National Institute for Food and Agriculture