REACCH
Summary Data Management Plan
2012

UNIVERSITY OF IDAHO
REGIONAL APPROACHES TO CLIMATE CHANGE (REACCH)

Updated as of: September 2012

Executive Summary

The following REACCH data management plan is a comprehensive strategic and planning document that outlines the steps for building and maintaining an environmentally-focused data analysis system. REACCH, which stands for Regional Approaches to Climate Change – is a research team composed of over 40 scientists, dedicated to understanding how climate alterations are affecting our Pacific Northwest agriculture.

The REACCH Data Management Plan has four key goals that are

- To develop an extensible and decentralized data management framework for climate based research;
- To promote the decentralization of data storage by instantiating all data access and rules in a programmatic application layer environment;
- Allowing researchers and stakeholders to analyze and examine climate based scientific research thru the use of easily accessible open source compiled and web based data tools
- Provide a library of climate based tools that allow the user to analyze and examine data from multiple sources –with no need to download any data.
- To develop a data policy framework for organizing and structuring how data is managed.

Introduction

What is Data Management?

"Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise."

Data management is a changing and dynamic discipline that has evolved extensively over time. With the development of personal computers – and the advancement in the use of database technologies, web development, and data communication protocols (XML) – data management has become an essential component to any large, diverse, data intensive project (citation, citation, citation)

Data management in scientific research has had considerable growth – with many funding agencies recognizing the importance of data management as part of an overall research strategy (citation). For example, both the National Science Foundation (NSF) and the National Institute for Health (NIH) have data management plan templates – and encourage each research proposal to have a data management plan.

The REACCH program is a prime example of a diverse project that requires data management structure, guidelines, and policy. With over 50 scientists, staff, policymakers, and hundreds of stakeholders – clearly describing and understanding our data management vision, plan, and policy is essential for success.

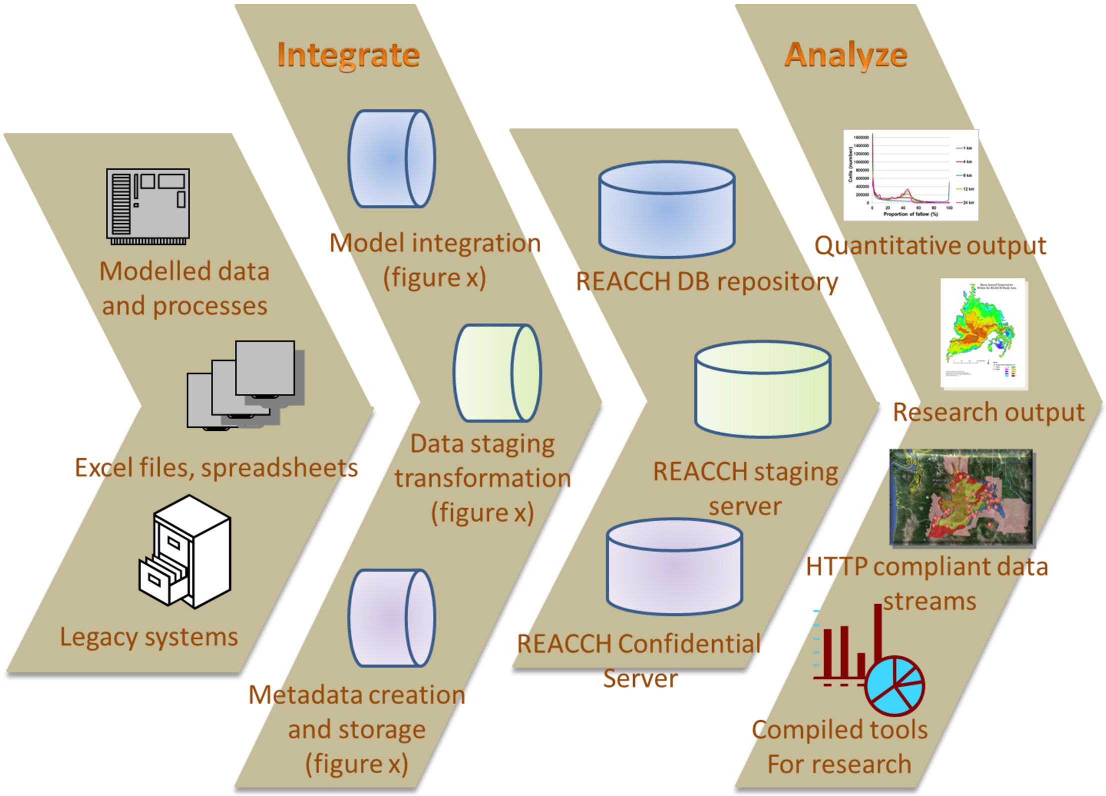
Integrate
Analyze
Modelled data and processes
Excel files, spreadsheets
Legacy systems
Model integration (figure x)
Data staging transformation (figure x)
Metadata creation and storage (figure x)
REACCH DB repository
REACCH staging server
REACCH Confidential Server
Quantitative output
Research output
HTTP compliant data streams
Compiled tools For research

Figure 1. REACCH Data Management flow

REACCH Data Management Vision and Strategy

REACCH's data management vision is focused on three aspects:

1. Modularity
2. Flexibility
3. Extensibility

Modularity. REACCH's scientific efforts are diverse. With scientists from areas including cropping systems, soils, biotics, education, biological engineering, atmospheric science, climatology, geography, and the like – a modular data approach indicates an effort that is somewhat standardized – and can be used interchangeably amongst groups. Metadata cataloging is an example area where modularity can be helpful in reducing time and effort – by developing a modular structure that can be stamped out again and again with alterations.

Flexibility. This is a changing, moving project. Scientists are exploring ideas and coming up with new and innovative ways to better understand our climate in the northwest. From a data management perspective, we want to be flexible in terms of how we develop our processes for data uploading, how we expose data, what applications we create, etc. The data management plan's purpose is to lay down a foundation for data management development – but to additionally be flexible enough to change as necessary when information or circumstances change.

***Extensibility*.** To be "extensible" means take into consideration the fact that REACCH will change and grow over time. We want our data management effort, our data structure, our metadata, to be formulated in a way that ensures that it remains extensible.

REACCH Data Management Goals

The REACCH data management goals –and associated objectives – are a high level view of our strategic data management direction. Our hope is that there is a direct association between:

- REACCH overall program goals;
- REACCH data management goals;
- REACCH data management objectives; and
- REACCH data management tasks as part of implementation and sustainability.

Goals:

1. *To develop an extensible and decentralized data management framework for climate based research.*

2. *To promote the decentralization of data storage by instantiating all data access and rules in a programmatic application layer environment.*

3. *Allowing researchers and stakeholders to analyze and examine climate based scientific research thru the use of easily accessible open source compiled and web based data tools.*

4. *Provide a library of climate based tools and information that allow the user to analyze and examine data from multiple sources –with no need to download data.*

5. *To develop a data policy framework for organizing and structuring how data is managed, that is clear to stakeholders, researchers, and the public.*

REACCH Operational Data Management Guidelines

Introduction

Development of a robust data management effort requires guidelines on how data will be used and managed over time. Particularly in an environment where data is being continuously added, analyzed, removed, and changed – there needs to be a set of processes and rules that define how this information will be managed – not just during the grant time period, but after the conclusion of the grant.
These operational guidelines describe steps necessary for operational management, and refer to additional informational collection documents that need to be in place to record said information.

Data Inclusion

Data inclusion guidelines define which datasets will be incorporated into the REACCH repositories. From a broad perspective, REACCH datasets can be grouped into the following areas:

1. Raw unprocessed data
2. Processed data
3. Analyzed data
4. Published papers and presentations

These functional groupings of data show how the variety of datasets and information being collected are related in several meaningful ways (Figure 2).

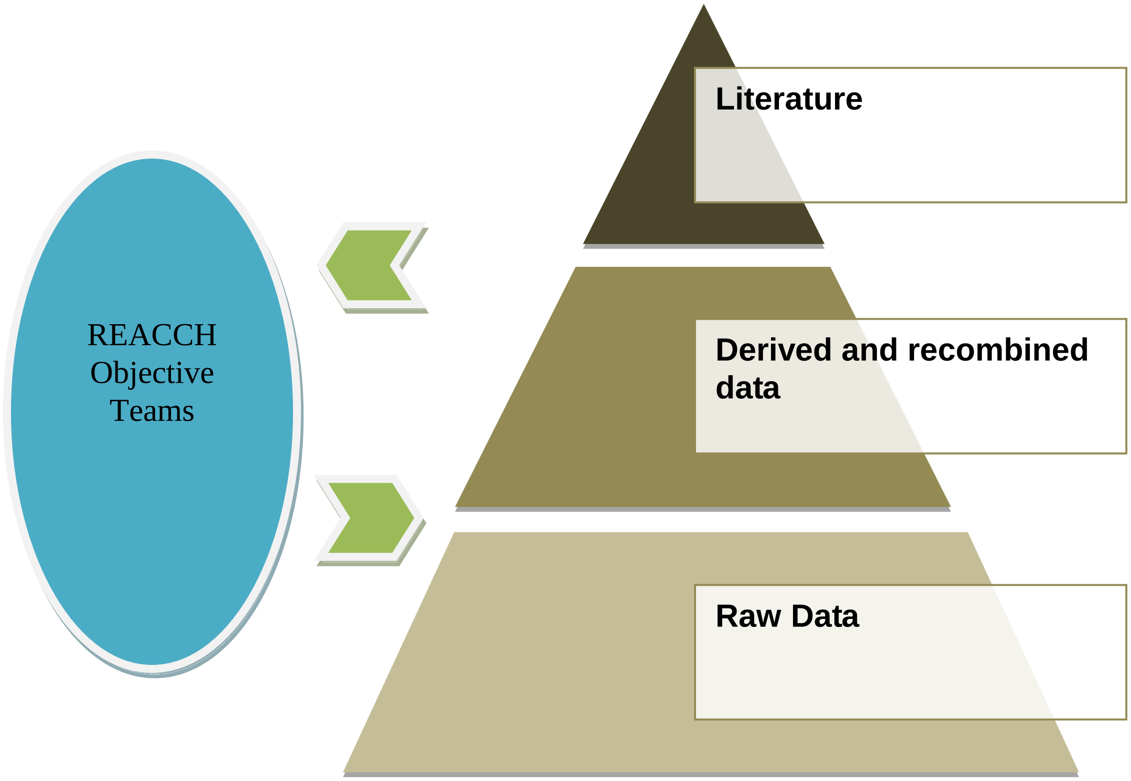

Figure 2: REACCH Semantic overview

Data retention

The retention of REACCH data involves the grouping of datasets into retention levels, as described below:

Retained indefinitely	Level1
Retained til end of project (YR5)	Level2
Retained yearly	Level3
Retained monthly	Level4.

The determination of retention level per dataset will be set by the owner of the data, and will be included in the appropriate metadata for the dataset. This information will then be used to segregate datasets when backup and redundancy components are put in place.

Data formats and dissemination

REACCH has a very diverse scientific grouping, with researchers in cropping systems, biotics, modeling, economics, and sociology. Several data input and output formats will be utilized, including:

Input formats

- Netcdf
- ESRI ArcGIS shapefile
- ESRI ArcGIS geodatabase
- Excel spreadsheet

Output formats

- ESRI based map service
- ESRI ArcGIS geodatabase
- CSV
- XML
- WMS/WFS map service
- THREDDS/OPENDAP protocol service

Dissemination

Dissemination will focus on the dynamic distribution and analysis of datasets thru web services – rather than data download and individual analysis. Given the complexity and size of datasets, the actual acquisition of a full dataset for analysis of an individual data type or component would not be efficient or appropriate. As such, REACCH's dissemination model will focus on the use of web server technology to allow researchers, stakeholders, and the public – regardless of the technology device (computer, mobile phone) to refine data output thru information query – and then allow said user to view the results of this query – or then download just the information that is being requested.

Data storage and preservation of access

REACCH data will be stored within a redundant geospatial database structure (PostgresQL /Linux), and managed thru web service protocols (using ESRI's geoportal server for data discovery and searching).

Data will be preserved thru primary and incremental backups – as well as replicated to other mirrored servers attached to the University of Idaho network.

REACCH Data Management Organizational Framework

REACCH Data Management Organizational Model

The REACCH data management organizational model is important to describe the flow of decision-making and for clarity in terms of how decisions regarding data management should be made.

- Evaluation of organization’s goals and objectives, and how they fit with the proposed initiative;
- Description of the decision-making work flow for the program
- Defining of roles and responsibilities for the initiative within the organization;
- Clear methods that describe how management can modify/re-evaluate approaches that are not working; and
- Well-defined mechanisms for communicating between participating groups.

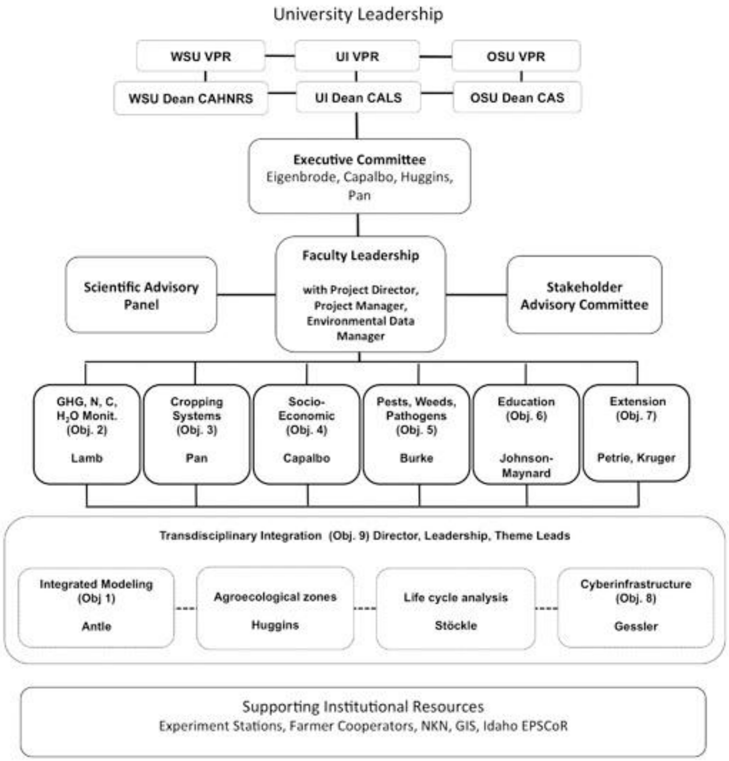
University Leadership
WSU VPR
UI VPR
OSU VPR
WSU Dean CAHNRS
UI Dean CALS
OSU Dean CAS
Executive Committee
Eigenbrode, Capalbo, Huggins, Pan
Scientific Advisory Panel
Faculty Leadership
with Project Director, Project Manager, Environmental Data Manager
Stakeholder Advisory Committee
GHG, N, C, H_2O Monit. (Obj. 2)
Lamb
Cropping Systems (Obj. 3)
Pan
Socio-Economic (Obj. 4)
Capalbo
Pests, Weeds, Pathogens (Obj. 5)
Burke
Education (Obj. 6)
Johnson-Maynard
Extension (Obj. 7)
Petrie, Kruger
Transdisciplinary Integration (Obj. 9) Director, Leadership, Theme Leads
Integrated Modeling (Obj 1)
Antle
Agroecological zones
Huggins
Life cycle analysis
Stöckle
Cyberinfrastructure (Obj. 8)
Gessler
Supporting Institutional Resources
Experiment Stations, Farmer Cooperators, NKN, GIS, Idaho EPSCoR

Figure 3: REACCH overall organizational model

REACCH Data Management Team Structure

The REACCH Objective 8 Infrastructure and Cyberinfrastructure team (Figure 3) will serve as the lead group coordinating data management, with direction from the Environmental Data Manager.

Each of the REACCH objective teams will work in coordination with Objective team 8 – as each team will have assigned tasks that are part of the overall REACCH Data Management Project Plan – and those tasks will be assigned in REACCH's intranet collaboration portal system - Central Desktop.

Data Management Data Flow

The REACCH approach (Figure 4) to the flow and management of data will utilize several key components:

- A internal collaboration web portal for informal information sharing, task and milestone management, and other documentation sharing;

- An external web site (www.reacchpna.org) for public information sharing and secure access to the REACCH data portal;

- A database repository (Linux operating system, utilizing a PostgresQL enterprise geospatial database) for data storage;

- An application web server model for data analysis, discovery, and upload/download;

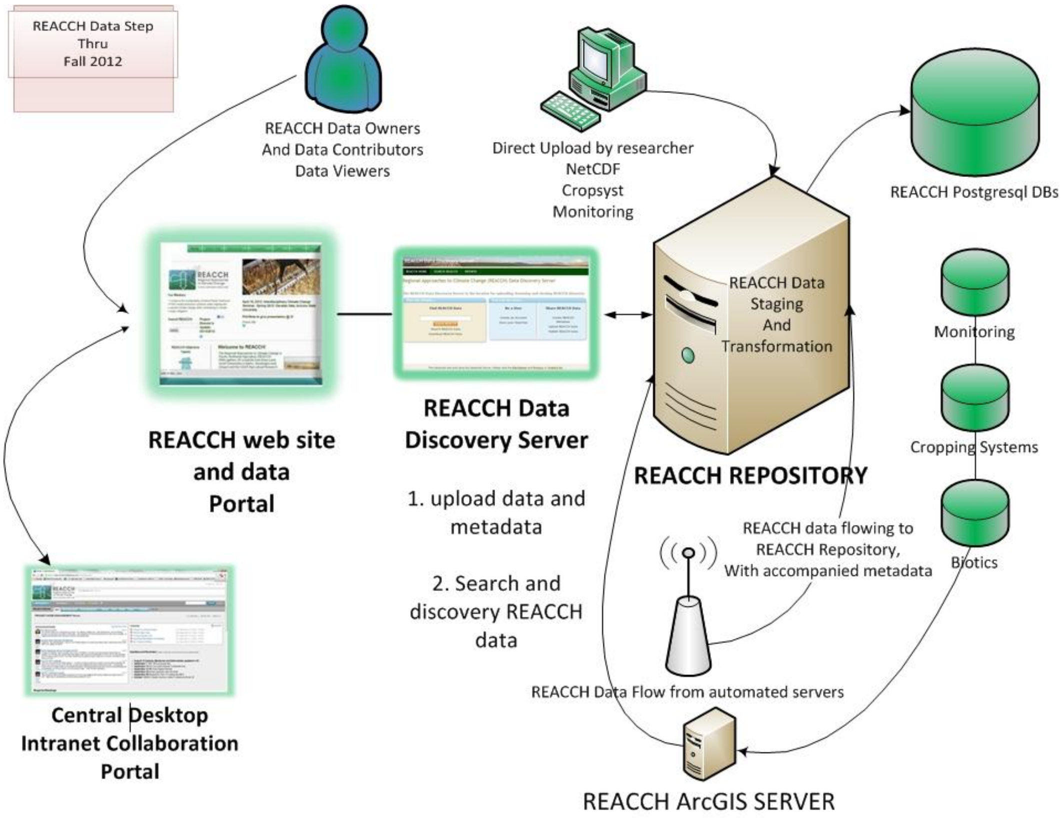

Figure 4: REACCH Physical Architecture

REACCH Data Management Roles and Policy

Roles

From a REACCH overarching perspective, an overall organizational framework is needed to identify each objective team’s role in the data management effort.

Some common precepts that help to outline this framework:

1. Data ownership determines decision-making regarding distribution
2. Datasets can have multiple roles (Data Creator, Data Owner)
3. Objective team roles determine their REACCH data management procedures over time.

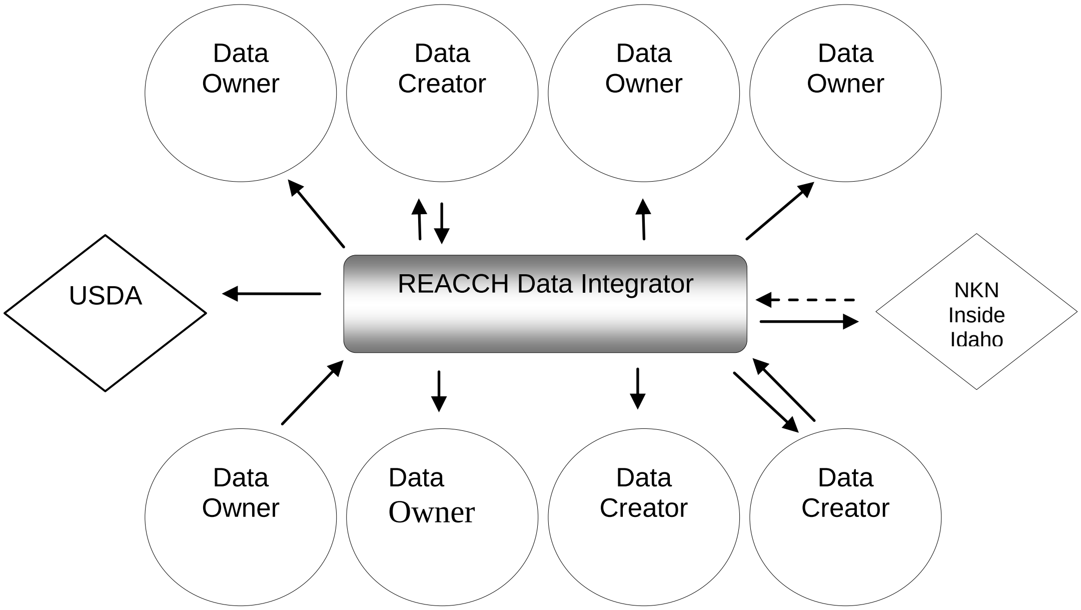

Figure 5: REACCH Data Organizational Framework Roles

Data Management Policy and Access Standards

REACCH has developed a Data Management Access Policy, that outlines guidelines for external data distribution, data inclusion, and the delineation of data based on issues related to privacy and confidentiality. In addition, the REACCH Data Management Access Policy provides guidelines to REACCH researchers on:

- How data will be segregated based on data type;
- How such information will be stored and managed;
- Guidelines for REACCH researchers on the methods of data uploading; and
- Potential barriers to data inclusion and how to resolve.

In addition to the REACCH Data Management Policy, REACCH data agreements have been developed for both general data access, and restricted data access – that will be presented to all users of REACCH data – via a web form – to ensure that essential information regarding the potential use and distribution of REACCH data is communicated. All users of REACCH data will be required to acknowledge review of these data agreements, based on the type of data that is to be reviewed/.analyzed/downloaded.

Metadata policy standards

What is metadata?

Metadata is descriptive information for your research data. Metadata describes said information, so we can search for this data, and find it easily using technology tools.

Metadata is a collection descriptive fields, or columns of data, that define core information about a dataset that helps someone who is unfamiliar with the information to know:

- Who created it;
- Where it came from;
- How it was created;
- When it was created; and
- What particular components of the dataset mean.

Why is metadata important?

Metadata not only helps to describe the data but helps the person examining it to know where it might be found - which is often times referred to as 'discovery'. Discovery metadata is a core aspect of your dataset.

REACCH is using a customized version of the ISO 19115 metadata standard.

"ISO is a network of the national standards institutes of 157 countries, on the basis of one member per country, with a Central Secretariat in Geneva, Switzerland, that coordinates the system.

ISO is a non-governmental organization: its members are not, as is the case in the United Nations system, delegations of national governments. Nevertheless, ISO occupies a special position between the public and private sectors. This is because, on the one hand, many of its member institutes are part of the governmental structure of their countries, or are mandated by their government. On the other hand, other members have their roots uniquely in the private sector, having been set up by national partnerships of industry associations.

Therefore, ISO is able to act as a bridging organization in which a consensus can be reached on solutions that meet both the requirements of business and the broader needs of society, such as the needs of stakeholder groups like consumers and users."[1]

ISO 19115 and 19115-2 define the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data

Standard metadata frameworks supported by REACCH

As noted above, REACCH will utilize the ISO-19115 metadata standard as the core standard for data storage. Given that the REACCH effort has a strong geographic data focus – the ISO19115 standard provides the most broad and extensive metadata element structure for storage of a wide grouping of data types.

In addition, the ISO standard provides strong crosswalks to other metadata standards – including EML, Dublin Core, Darwin Core, and other widely used metadata standards.

[1] International Standardization of Security - http://www.iso.org/iso/iss_home

REACCH Data Management Proposed Design and Architecture

Design Overview

The design of the REACCH data management effort is focused on three core areas:

- Development of a strong database repository for storage of data;
- Use of a data discovery server for the uploading and tagging of data – as well as for searching, or ‘discovering’ REACCH data, and
- The use of web application server technology (ArcGIS Server, THREDDS) to organize and display data in a web services protocol – for analysis, download, and exposure of said data to other systems and repositories.

The above design is an approach that is used for the collection and distribution of large, diverse datasets, particularly those that may be very large, cumbersome, and detailed in nature. To simply provide said datasets for download is, to say the least, inefficient.

The REACCH model is an attempt to organize and distribute data in a way that permits the user to organize how they wish to see the data, and allow server-side technology to process and select the requested information – and present it for analysis, and specific download, if that is needed.

Conceptual/Semantic Design

The conceptual design of the REACCH data management model is seen in Figure 6 below.

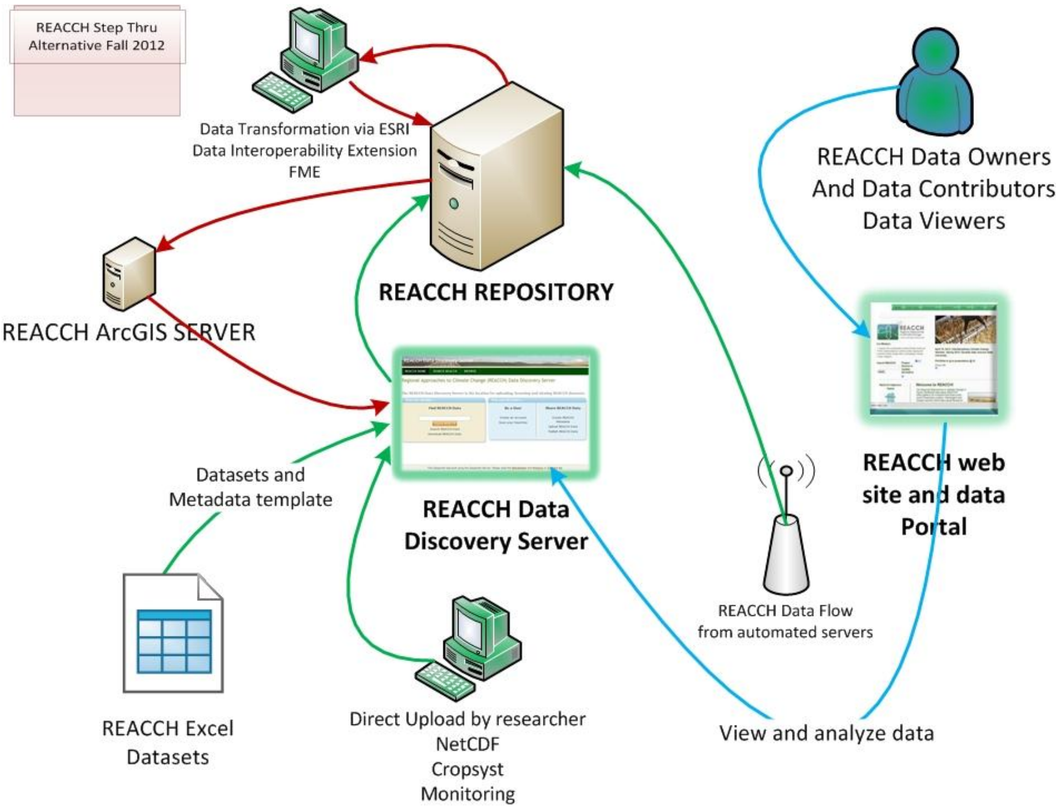

Figure 6: REACCH Conceptual Data Design

REACCH data providers will enter the system thru the REACCH web site – where, based on their need, they will be directed to the REACH data discovery server. The REACCH data discovery server will be the location where users will upload their datasets, and associated metadata. This data will be deposited into the REACCH repository (Linux/PostgresQL), where at that point, a set of transformative processes will be run – and placing the dataset into the appropriate location within the REACCH geospatial database.

After review, approval, and publishing of loaded data - REACCH application servers – with specific applications developed for functional information distribution (biotics, cropping systems, economics, education, modeling, etc) will be developed – and will access the REACCH geospatial database – displaying data to users for search and discovery thru the REACCH data discovery server.

Logical/Physical Design

REACCH is leveraging the University of Idaho's Northwest Knowledge Network (NKN - www.northwestknowledge.net) as a source for infrastructure and networking support. NKN is providing server and networking systems that allow REACCH to implement the conceptual design indicated in Figure 6.

REACCH's physical design utilizes virtual server technology to build out our database repository, application servers, web server, and our www.reacchpna.org web site. All systems will additionally be redundantly mirrored at other locations utilizing the Idaho IRON high speed network.[2]

[2] Idaho Regional Optical Network, http://www.ironforidaho.net/