DOE Final Report

DOE award number: DE-SC0005868, awarded to UAF

Alaska Climate Data Center, PI Larry Hinzman

12/6/2013

Comparison of actual accomplishments with goals and objectives established, and why established goals were not met:

Objectives	Actual Accomplishments
• Serve as a hub for diverse data sets	GeoNetwork and DSpace
• Develop a central web-based "gateway" to easily access Alaska/Arctic projects	http://data.alaska.edu
• Conduct research into climate and weather impacts facing the state	Compiled extensive data holdings on environment/climate dynamics, computational modeling w/ WRF, CCSM, PISM, others; downscaling of climatological data.
• Semantic infrastructure	significant progress, detail below
• Broaden modeling and storage capabilities	64-core servers w/77 TB storage

A discussion of what was accomplished under these goals.

A tremendous amount has been accomplished since the project inception, and we are quite satisfied with the data center that was created, essentially starting from nothing. This center is already garnering a national and international reputation due to the value of the holdings, the ease of archiving and accessing the data and the accepted security of the system. The focus of Arctic data documenting the environmental response to a changing climate offers the opportunity for enhanced analysis and discovery as previously unknown data streams are made openly available for public access.

GeoNetwork

The IARC Data Archive (IDA - http://climate.iarc.uaf.edu/geonetwork) is the first major software component of the Data Center. IDA is designed for long-term archival storage of Arctic data assists, and has been recognized by the NSF as an official Archive. IDA runs a customized instance of GeoNetwork (http://geonetwork-opensource.org/), a multi-lingual interface providing:

- Full search of local and distributed geospatial catalogs.
- Ingestion and fine-grain access control for data and graphics.
- An interactive Web Map Viewer to combine distributed Web Map Services.
- An online metadata editor, with templates for ISO 19115 and others.
- Harvesting of metadata from other catalogs.
- Support for OGC-CSW 2.0.2, OAI-PMH, and Z39.50 protocols.
- An RSS feed.

The GeoNetwork instance is hosted by a xen virtual infrastructure, currently consisting of two 16-core servers attached to two ISCSI disk arrays of 30 TB each, and a tape-silo server with 20 TB of disk and 240 TB of LTO-5 tape. Additional offsite storage is provided by the Arctic Region Supercomputing Center.

Current holdings consist of 177 data sets for a total of 4.5 TB, not counting replicates. Holding are expected to rise significantly in the near future.

A data policy was drafted and approved by the user community, see
http://data.iarc.uaf.edu/docs/DataHandlingPolicy2011.html
This data policy was vetted with Japanese collaborators to ensure compatibility with numerous collaborative research projects and to ensure continued free and open access to multiple data streams.

Semantic Infrastructure

Pieces constructed to date on this project, as part of a graduate degree in Computer Science:

- a parser to turn ISO 19115 XML metadata into RDF statements.
- an ontology describing relationships amongst parse-generated RDF.
- an RDF data store to hold parse-generated RDF, along with RDF inferred by the ontology.
- a web application tying the above together, with SPARQL querying capability of the RDF store.

Future efforts will concentrate on generating RDF statements that describe the physical infrastructure, to be ultimately combined with management policy ontologies for both hardware and data. It is anticipated that the semantic management infrastructure will also lay the groundwork for data integration efforts. Ideas relevant to semantic goal-oriented methods of combing integrated data with software into workflows will form the basis of future proposals.

Broaden Modeling and Storage Capabilities

Four 64-core servers came online in 2012 at IARC, used for jobs that don't fit into the normal queue structure at ARSC (next section). A 77 TB replicated storage system (154 TB total) came online at IARC in 2013, accessible by a wide variety of clients, both virtual and physical. Also in 2013, LTO-5 tape drives were augmented with larger capacity LTO-6 drives in the tape silos.

ARSC Data Infrastructure

The Arctic Region Supercomputing Center (ARSC) has designed, implemented, and put into production a new infrastructure for data hosting. Offered services are open to any University constituent, and are described in some detail online at http://www.arsc.edu/arsc/resources/web-based-services/index.xml. This infrastructure is built upon the large-scale storage resources managed by ARSC, which now include over 800TB of disk and a tape silo with 30PB capacity.

The new infrastructure utilizes virtual machines on a high-availability Linux cluster. The VMs are scaled as needed for different data projects, and given access to needed CPU, memory and centralized storage as needed. Some data projects, such as http://weather.arsc.edu, make use of the hierarchical storage manager offered by ARSC (weather.arsc.edu is ARSC's largest current data portal, with over 270TB of weather forecast products). Others, such as http://data.alaska.edu, instead rely dedicated high-performance storage (i.e., Fiber channel arrays of disk drives). Services on the VMs may utilize MySQL or PostgreSQL database servers, which are on a separate enterprise-class database server.

The services deployed on VMs are monitored 24x7, backed up, and supported by ARSC's help desk (with partners, such as the Rasmuson Library and OIT).

ARSC/UAF Common Authentication

The University of Alaska has a central authentication system and enterprise directory. This is accessible via secure LDAP (LDAPS), Shibboleth, and Active Directory. In fall 2011, ARSC enabled this University-based central authentication system on all academic systems. This meant that the same username and password utilized for other campus services would work for ARSC services. This replaced a system that required every ARSC user to have a username/password that was separate from the UA username/password. The old system was removed in January 2012, resulting in easier utilization by ARSC users and greater use of shared central campus resources and support.

As part of a Master of Science in Computer Science, an ARSC graduate research assistant deployed DSpace with the same centralized common authentication. By deploying LDAPS along with DSpace, the result is a mechanism that automatically creates a DSpace username for anyone authenticated and authorized by the campus system. While anonymous users are able to view some data on DSpace, a basic set of roles is automatically granted upon authentication, so that data restricted to campus use may also be visible. The automatically-created username may be given additional roles by individual communities (such as the Library, or the Graduate School).

Schedule Status - milestones, anticipated and actual completion dates

The Archive portion of the Data Center is the first major milestone, fully functional as of November 2011. Work on the second major milestone is ongoing, a semantic management infrastructure to cover hardware associated with storage devices. The third major milestone is a broadened IARC modeling and virtual infrastructure storage, successfully completed during 2012 and 2013. Beyond 2013, we anticipate continued exploration of semantic and virtual infrastructures to facilitate data integration and goal-oriented workflow generation.

Any changes in approach or aims, and reasons

The objectives and plans are being implemented directly as originally planned. It was necessary to focus more of the funding support to programmers and archivists (as opposed to research scientists) than originally planned because developing the core archive was somewhat more demanding than originally expected. It was obvious early on that the data archive must be fully functional before data streams may be dynamically incorporated.

Actual or anticipated problems or delays

The data center was implemented as originally planned and described. A no cost extension was requested, as it took longer than originally anticipated to become fully staffed. The expertise required to implement semantic technology into a functional data center are actually both rare and highly in demand.   Consequently, it took longer than expected to locate and hire the staff needed to implement these plans. Work is ongoing. A second no cost extension was requested to broaden storage capability.

Changes in key personnel

A post-doc was not required to write ontologies, as a gifted undergraduate mathematics student successfully executed a large portion of that work. In early 2013, our Data Archivist accepted another position.

Deliverables, Collaborations

- Data Archive established in Feb. 2011, and fully functional by Nov. 2011
  - http://climate.iarc.uaf.edu/geonetwork
  - Added Japanese language support in 2012
  - xen virtual infrastructure
- An MS degree in Computer Science (summer 2012), detailing semantic technologies as used in the Alaska Climate Data Center
  - web app as described in "Semantic Management Infrastructure"
- An MS degree in Computer Science (May 2012), implementing DSpace and deploying single source authentication
  - http://dspace.alaska.edu
- Four 64-core servers came online (2012)
- 77 TB of replicated storage (154 TB total) came online (2013)
- Collaboration with Arctic Region Supercomputing Center (ARSC)
  - http://www.arsc.edu/arsc
  - Launch of Alaska Data Central (2013), http://data.alaska.edu
- Collaboration with Geographic Information Network of Alaska (GINA)
  - http://www.gina.alaska.edu
- Collaboration with Japan Aerospace Exploration Agency (JAXA)
  - http://www.ijis.iarc.uaf.edu
- Collaboration with UAF Rasmuson Library