COL/CE/10

Intelligent Data: Applying RSS to Scientific Data Resources

FY 2003 Proposal to the NOAA HPCC Program

August 19, 2002

| Title Page | Proposed Project | Budget Page |

Principal Investigator: **Ann Keane**

Line Organization: OAR
Routing Code: R/ETL
Address: Environmental Technology Laboratory
325 Broadway
R/ETL
Boulder, CO 80305

Phone: (303) 497-7424
Fax: (303) 497-6978
E-mail Address: Ann.Keane@noaa.gov

Proposal Theme: **Technologies for Collaboration, Visualization, or Analysis**

Funding Summary: FY 2003 $ 31,400

Ann Keane
Computer Specialist
Environmental Technology Laboratory

Richard H. Beeler
Chief, Computing and Network
Environmental Technology Laboratory

William D. Neff
Director
Environmental Technology Laboratory

Intelligent Data: Applying RSS to Scientific Data Resources

Proposal for FY 2003 HPCC Funding

Prepared by: Ann Keane

Executive Summary:

With the rapid adoption and growth of the World Wide Web (WWW) as a means of disseminating research data, little has been done to make finding and assimilating data easier under realtime, operational conditions and over low-bandwidth devices. A parallel to this problem has been solved by WWW news sites seeking to deliver the latest headlines, stock quotes and sports scores to wireless clients such as cell phone and PDA's. The development and widespread application of the RSS (Rich Site Summary) standard allows news sites to publish their latest information and distribute it to be assimilated by other sites. This process is carried out automatically by computer programs or clients. By characterizing research data with RSS, it is possible for scientists to configure a client to automatically assess available data, acquire it and make it available to them anywhere in the world. Adapting RSS technology to research, scientists and a number of other users gain the ability to create specialized, up-to-the-minute data collections from resources spread across the internet and access them remotely through lightweight clients.

Problem Statement:

An overwhelming amount of information is available to researchers over the WWW and is distributed through a number of sites which does not allow for timely or comprehensive consultation during field programs where network bandwidth may be limited.

Background

In 2000, a group of researchers using the NOAA P-3 research aircraft received a grant from HPCC to equip that aircraft with a satellite transmitter-receiver and develop a prototype wide-area network (WAN) for research aircraft. As part of the demonstration of the WAN's capabilities, data were sent to researchers on the aircraft for the direction of the experimental flights and data were sent by the aircraft to the ground, for use by NWS forecasters. Given limited bandwidth and the broad range of possible data requests, two technicians acted as gatekeepers, one on the ground and one on the aircraft, to prioritize and regulate the exchange of information between the two. The demonstration of WLAN technology was successful. Satellite communications technology was added to both NOAA P-3's, substantially increasing the available bandwidth.

In 2001, the PACJET program again used satellite communications aboard the NOAA P-3 for two-way communications, this time relying on an automated ground station to collect data of interest to the aircraft researchers. A single gatekeeper on the aircraft selected data for upload and responded to data requests from the ground. Given increased bandwidth, the upload of a single satellite image was reduced from 10 minutes to 2 minutes. The gatekeeper was still required to assess the available data and prioritize requests. At times when specific data sources were not available the gatekeeper would search for others over the WWW.

These projects revealed several remaining problems which anyone developing data services for remote, light-weight clients (such as cell phones or personal digital assistant's (PDA's)) will encounter; the need for the user to actively participate in the data exchange, the number of diverse resources which may be required and may not be available, the inability of advertising new or replacement resources and the large bandwidth needed to access most research data resources.

Application to HPCC Program Objectives

This problem is central to HPCC's goal to provide better access to the real-time data generated by NOAA. As internet equipped cell phones and PDA's become more common, there will be a growing need to address their specific needs and limitations. Successful demonstration of automated metadata generation, data aggregation and data dissemination technologies can improve NOAA's mission response and customize large volumes of data to specific user needs.

Proposed Solution:

RDF Site Summary or Rich Site Summary (RSS) is a light weight XML format developed to describe web content in a universal, extensible format. It is most widely used as a means of distributing the headlines and news content of websites to clients on other websites which can automatically incorporate the remote content on the local site. It is also used to request sports scores and stock prices in near-realtime. Originally developed by Netscape as a means of supplying information for the My Netscape portal, RSS is now an open standard which is XML and RDF compliant. The standard is available at http://www.purl.org/rss/. For an overview of RSS development: http://www.xml.com/pub/a/2000/07/17/syndication/rss.html. A working example is http://www.oreillynet.com/meerkat/.

To solve the specific problem of providing automated field support data, this proposal will develop a server to generate, acquire and aggregate RSS metadata for research data resources. This server will provide a web form for scientists to select the data resources in advance of a field campaign. The server will then query these sources, cache available data products and write RSS metadata for that product and aggregate metadata for the whole collection. This RSS metadata would then be exchanged over a light client, informing the remote scientist of available resources. Scientists in the field would then able to upload the latest data either through standard means (WWW or FTP) or through a personalized webpage. This data collection would be available to collaborators anywhere in the world. Data from remote platforms could also be advertised and requested through an exchange of RSS metadata. In the case of missing data, backup source would be made available. Scientists will also be able to program the client to retrieve data automatically, decreasing the amount of time they have to spend acquiring the data.

To make RSS available to the wider NOAA community, this proposal will also create a web-based RSS tutorial and engine to generate RSS for NOAA websites and products.

Analysis:

For many applications there is a finite time window in which data provides useful insight. In operational environments, that data must be available when its needed and presented in a way that is easily absorbed and acted on. Any solution must allow diverse, distributed data resources to advertise their products in a low bandwidth manner and allow the user to select and acquire products with minimal interaction. Configuring data such that a computer client may act on it for the user will decrease the amount of time the user takes to acquire the data and increase the amount of time available for assimilating its meaning. Adopting an extensible, standards based solution increases the likelihood of broad adoption and maintainability over time.

RSS is a well established and widely used technology for the distribution and assimilation of realtime information which lends itself well to the operational environment. Given its XML foundation, it can be universally exchanged and extended as data services are developed. It can be used as a stepping stone toward more comprehensive XML web services.

Alternatives

As stated in the problem, prior solutions to this problem have included:

1. Human interaction and exchange, the many drawbacks of which include, expense, weight, and the need to sleep.

2. Automated scripting developed for specific use is able to acquire information automatically, but limits the amount of information available to the user and the ability of computer clients to interact with the data.

A third alternative developed in conjunction with number 2, would be to create and transmit a webpage advertising the contents of the acquired data archive. This would again limit a client's access to the data and increase the amount of bandwidth used to advertise data resources.

Performance Measures:

- Demonstrate the usage of NOAA data through RSS by researchers, forecasters, and the general public.
- Increase the number of NOAA resources available through RSS.
- Assess the use of this technology for accessing NOAA data through wireless devices.

Milestones

Month 1 – Develop demonstration aggregation server and client, deploy for one field season.
Month 2 – Develop tutorial materials, monitor usage and feedback, final report.

Deliverables

- A demonstration web-based aggregation server and client.
- A means of generating and distributing RSS metadata for NOAA webpages and data sources.
- A tutorial for NOAA and collaborators to generate RSS metadata for aggregation sources.
- Project final report