An Overview of EPA's System of Registries (SoR)

Michael Pendleton
U.S. EPA, 1200 Pennsylvania Ave, N.W., Washington D.C. 20460
pendleton.michael@epa.gov

and

Ky Ostergaard
Science Applications International Corporation
6565 Arlington Blvd, Suite 101, Falls Church, VA 22042
ostergaardk@sdc-moses.com

ABSTRACT

The Environmental Protection Agency's (EPA) System of Registries (SoR) is the product of a decade long effort to better manage Agency information resources in order to support environmentally sound decision-making. The SoR provides a gateway and search capability to several registries and repositories residing in EPA's Office of Environmental Information (OEI). The registries contain identification information for objects of interest to EPA, Network trading partners, including states and tribal entities, and the public. These objects consist of data elements, XML tags, data standards, substances (chemicals, biological organisms, and physical properties), terms, facilities, regulations, and data sets that the Agency uses in its core business processes.

These registries comprise a critical link in EPA's information architecture and are a vital component to the National Environmental Information Exchange Network (Network). Specifically, the SoR was developed to support the Agency's data standards program and numerous Agency information technology initiatives, including the Agency architecture and data exchange with stakeholders through network nodes.

The presentation accompanying this paper provides an overview of the information resources being offered through the System of Registries website, which is located at www.epa.gov/sor.

History

The System of Registries represents a 10-year effort to manage Agency information resources as strategic assets and an increasing recognition that environmentally sound decision-making must be based on accurate data and information. This effort was initially conceptualized in the early 1990's in response to numerous Office of Management and Budget, General Accounting Office and Inspector General documents identifying the need to improve longstanding information technology weaknesses. At the time, the Agency was described as being "data rich but information poor."

In response to these reports, EPA's Office of Information Resources initiated an agency-wide, strategic information management planning process that included an Information and Data Management program focused on establishing data management policies and standards to improve and maintain data integrity.

From a concept paper in 1993 calling for an incremental approach to building a system of registries and data standards, to the initial implementation of the Environmental Data Registry in 1997, promoting metadata management has been challenging. However, senior EPA management demonstrated their commitment to metadata management in 1997 when EPA's Administrator announced the Reinventing Environmental Information initiative, which called for establishing data standards in a publicly accessible metadata registry. The development of the registry's capabilities through the years was done incrementally as funding became available and with significant interest and participation from the international standards community. In 2003, the System of Registries became an integral component to the Agency architectural plans.

EPA's Office of Environmental Information has sponsored the development of the System of Registries from its inception, with various partners, including Environmental Commissioners of the States (ECOS), the Environmental Data Standards Council, and numerous program offices within EPA.

Current Status

The System of Registries is a Web-based collection of metadata registries and repositories residing in the EPA's Office of Environmental Information. A registry is an official and authoritative list of specific, well-defined items of interest to an organization. The registries that comprise the SoR provide identification information for objects of interest to EPA, trading partners, including states and tribal entities, and the public. These described objects consist of data elements, XML tags, data standards, substances (chemicals and biological organisms), terms and definitions, facilities, regulations, and data sets that the Agency uses in its core business processes. These registries comprise a critical link in EPA's information architecture and are a vital component to the Exchange Network developed to facilitate data exchange with stakeholders through network nodes.

The registries that comprise the SoR are as follows:

1. Environmental Data Registry (EDR) – The EDR is a comprehensive, authoritative source of reference information (metadata) about environmental data including data elements, XML tags, and value domains. The EDR also supports the creation and implementation of data standards that are designed to promote the efficient sharing of environmental information among EPA, states, tribes, and other information trading partners.

2. Information Resource Registry System (IRRS) – The IRRS is an authoritative source of information about EPA applications, databases and other information resources. The IRRS will serve to integrate several information gathering

requirements and reduce the presently duplicative reporting burden for program offices and regions.

3. Substance Registry System (SRS) – The SRS serves as the nucleus for linking information about substances regulated by EPA. The SRS includes information on chemicals, organisms, and physical characteristics in EPA regulations, data systems, and other information resources.

4. Terminology Reference System (TRS) – The TRS provides a single resource of environmental terminology for EPA by compiling collections of terms and definitions from the Agency and other sources.

5. Exchange Network Registry – Currently considered an interim solution, the XML Registry for the Environmental Information Exchange Network will provide the capability to share information about XML Data Exchange Template (DETs), XML Schemas, Namespaces, WSDL files, and other supporting files needed to map data flows between partners. The Registry will contain information about schemas approved for use on the Network, as well as information about schemas under development. In time, the XML Registry will provide a clearinghouse for information related to data flows on the Network.

Virtual linkages from the SoR exist to:

6. The Facility Registry System (FRS) – a centrally-managed database that identifies facilities, sites or places subject to environmental regulations, or of environmental interest; and

7. The Environmental Information Management System (EIMS) – manages descriptive information on scientific data sets.

Content of the Metadata Registries

The registries are simply different views into a single, integrated metadata registry. This allows a user to locate a single regulation or application system and easily access the related terms, data elements, and substances.

The Environmental Data Registry contains 9,827 data elements. These data elements include those from 70 major information systems, as well as those from Agency data standards. Almost 86,000 substances from 956 information resources are contained in the Substance Registry System. Almost 11,000 terms are contained in the Terminology Registry System. Registry contents will continue to grow as new applications are registered.

Dynamic Linkages to Other Applications

The Registries include the Environmental Metadata Gateway (EMG) that serves as a stand-alone application that provides customized access into information in the System of Registries. It includes multiple search engine portals which publish information from the underlying SoR metadata in EPA default or user defined look-and-feel templates. The EMG enables users, both inside and outside of EPA, to search and seamlessly navigate to their pages displaying metadata registry content using a URL. In the future, EMG will expand with integrated search capabilities and the ability to transfer information to and from the registries in XML through the XML Gateway feature.

At this time, the EMG portal is being used to serve standard data element metadata to the Web site of the Environmental Data Standards Council and XML metadata to the Environmental Exchange Network XML Registry.

EPA's Office of Air and Radiation (OAR) partnered with OEI to develop the Radiation Information System for Cleanup Sites (RISCS) as a repository for publicly available, descriptive information on radiation cleanup sites. Radiation site cleanup information is collected by multiple programs and many different agencies. It is envisioned that a single repository of radiation site information will provide for better analysis of cleanup progress at a national level. A key feature of this application is its linkage to the System of Registries. Site description information for RISCS is provided from, and updated by, the Facility Registry System. Site contaminant identification information is provided from the Substance Registry System. RISCS application metadata will be stored in the Environmental Data Registry. RISCS will have dynamic links into all three applications to pull data and metadata to support the application.

Efforts are underway to link to the Envirofacts Information Warehouse to dynamically provide information on regulated substances to the various databases in the warehouse.

Uses of the Metadata Registries

Metadata in the registries is organized using the metamodel for metadata registries from ISO/IEC 11179, Part 3. Data elements in the EDR, terms in the TRS, and substances in the SRS are registered in association with an information resource, which might be an application system, a standard document, a thesaurus, or a law or regulation. The information resources are loaded into the data base as classification schemes, which enable hierarchical organization of information within schemes. In this way, the registries document systems, their structure, and the semantic meaning of their contents. When the metadata is used with data in a particular database, it helps users to understand the meaning of the data.

By combining system metadata from many applications in a single registry, the EDR can be used to analyze data across the organization. A specific set of tools have been developed to support database harmonization and integration. Harmonizing data improves the consistency and comparability of data across information systems, usually within a particular program area. The harmonization process identifies duplicate data storage or conflicting representations of the same information and provides the information needed to integrate data systems, or improve the consistency of data meaning and format across related systems. Harmonizing data can be achieved through a data mapping process.

Users can find EDR information useful in systems reengineering to conform to standards. The registry makes information available on standards that have been approved, standards under development, and the due dates for standards implementation. The standard data elements have been divided into subject area groups to assist design of database entities.

The EDR can assist system developers perform many tasks during their development process. It can help developers identify data standards to download and incorporate into their systems to enhance system interoperability and data exchange and comply with Agency policy. The EDR's search capabilities provide various methods of searching for metadata. The EDR enables developers to download data elements and attributes, allowable values (e.g., code sets), and other types of metadata to use in system design and documentation.

For the first time, the registries were recently used to support the development of a regulation, using data standards to shape information collection. The use of data standards to drive the regulatory life-cycle process will enhance the Agency's ability to integrate new information collections with other Agency information resources. EPA's Office of Wastewater Management requested support in establishing well-formed data requirements as part of a new Concentrated Animal Feeding Operations rule. Data standards team members met with the regulation writers to review existing requirements in the regulations, and provided a list of well-formed data elements, from the EDR, for inclusion in the regulation and on the Form 2B permit application.

Business Benefits

The System of Registries serves as the backbone to EPA’s data standards program. It not only serves to store standards metadata, but also provides public access to standards under development, and the ability for system developers to conduct harmonization of programmatic metadata prior to standards implementation and conformance. The registries serve to make well-formed data specification available for reuse. The System of Registries supports EPA’s enterprise architecture by providing access to the full spectrum of Agency metadata. By supporting the harmonization of disparate data, the registries help improve data consistency and reduce costly duplication associated with redundant storage and maintenance.

There is a plan to use the System of Registries to serve metadata for the Agency enterprise repository. The SoR is the agency’s premier tool in supporting public access to metadata resources. Monthly hits to the SoR Web site are typically about 70,000 requests, with .gov requests accounting for about 12 percent of the activity.

Future Activities

The System of Registries is currently being integrated into the enterprise architecture; a linkage with the enterprise repository is being built out and merging of XML metadata with data element metadata is underway. The System of Registries is currently being built out to serve as the host site for the Exchange Network. Ongoing documentation of existing applications will always be a priority as well using the tools to support conformance review and harmonization. A web-based, password protected system of access is being developed to support a program of data stewardship to be implemented by the Agency.

How can I find out more about the System of Registries?

System of Registries
www.epa.gov/sor

Environmental Data Registry
www.epa.gov/edr

Substance Registry System
www.epa.gov/srs

Terminology Reference System
www.epa.gov/trs

Information Resources Registry System
www.epa.gov/irrs

Facility Registry System
www.epa.gov/enviro/html/facility.html

Environmental Information Management System
www.epa.gov/eims

Exchange Network Web Site
www.exchangenetwork.net

EPA Representatives:

Larry Fitzwater, Mike Pendleton – System of Registries (General)
John Harman, Mike Pendleton – Information Resources Registry System
John Sykes – Environmental Information Management System
Pat Garvey – Facility Registry System