

Abstract
Federated Collections - Synergy Through Sharing
By Bob Donohue

The DOE research community requires easy and timely access to both current and legacy scientific and technical information in order to carry out the research and development missions of the Department. Using existing information to supplement and support their projects and programs, researchers create new information that expands the scientific and technical knowledge base available for future research activities. Through this cyclic phenomenon, science is advanced, breakthroughs are accomplished, and new technologies, processes, and products are developed. Critical to this process is the sharing of information in a consistent and organized manner to facilitate its reuse by the research community.

At present, the results of DOE-sponsored research are provided to the Department's Office of Scientific and Technical Information (OSTI) for organizing and announcing availability in compliance with DOE mandates; however, there is a great deal of data and information that does not meet the criteria for submission to OSTI that has significant potential value to researchers. This information and the full-text information submitted to OSTI generally resides in repositories at the originating sites, frequently supplemented with organized metadata and full text information residing on electronic databases for site-specific uses. Each organization builds these information systems to meet its own unique requirements, with little regard for the information needs of others who could potentially benefit from the information these systems contain. Information sharing that occurs now is largely performed through many ad hoc, point-to-point system interfaces.

With the migration from paper-based systems to an electronic environment, new opportunities are provided in the sharing and access of information that currently resides in these site-specific information systems through a distributed collections concept that allows for information searching, discovery, and retrieval across systems, greatly enhancing the availability and value of information in the research process.

The Federated Collections project is the proof-of-concept in terms of developing user options, exploring search efficiencies and information access and retrieval capabilities.

To ensure success, this rapid prototyping effort, in the form of a proof-of-concept project, requires the collective expertise and involvement of partners who share the common vision and goals of the STI Program. OSTI is actively seeking collaborators who are willing to participate on the Federated Collections rapid prototyping team.

Federated Collections - Synergy Through Sharing

The Department of Energy (DOE) research community requires easy and timely access to both current and legacy scientific and technical information (STI) in order to carry out the research and development missions of the Department. Using existing information to supplement and support their projects and programs, researchers create new information that expands the scientific and technical knowledge base available for future research activities. Through this cyclic phenomenon, science is advanced, breakthroughs are accomplished, and new technologies, processes, and products are developed. Critical to this process is the sharing of information in a consistent and organized manner to facilitate its reuse by the research community.

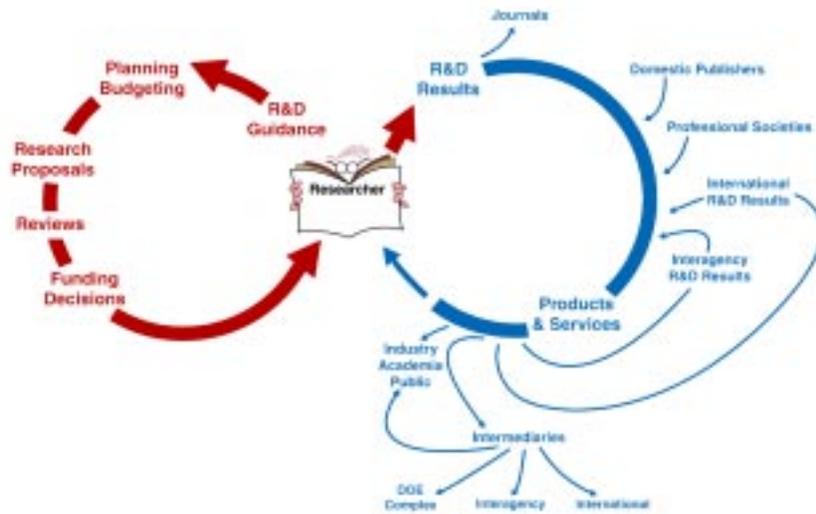


Figure 1 *STI Creation Cycle*

At present, the results of DOE-sponsored research are provided to the Department's Office of Scientific and Technical Information (OSTI) for organizing and announcing availability in compliance with DOE mandates; however, there is a great deal of data and information that does not meet the criteria for submission to OSTI that has significant potential value to researchers. This information and the full-text information submitted to OSTI generally resides in repositories at the originating sites, frequently supplemented with structured metadata and full text information residing on electronic databases for site-specific uses.

Federated Collections may be described on two levels: 1) on a technical level we are addressing multiple-source collections and federating (mapping and merging) them and searching the material as a single virtual collection - that is, one query searches across multiple repositories of STI where ever it may reside ; 2) on another level we are suggesting that each site or laboratory is part of a federation that

has similar interests in the accessibility and preservation of STI. Federation suggests collective sharing of responsibility, accountability and participation in these activities.

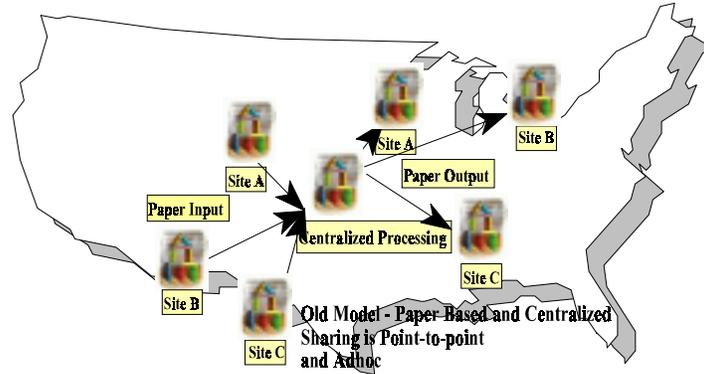


Figure 2 *The Old Model*

The Old Model

Each organization, understandably, builds these information systems to meet its own unique and parochial requirements, with little regard for the information needs of others who could potentially benefit from the information these systems contain. Information sharing that occurs now is largely performed through many ad hoc, point-to-point system interfaces.

The New Model

With the migration from paper-based systems to an electronic environment, new opportunities are provided in the sharing and access of information that currently resides in these site-specific information systems through a distributed collections concept that allows for information searching, discovery, and

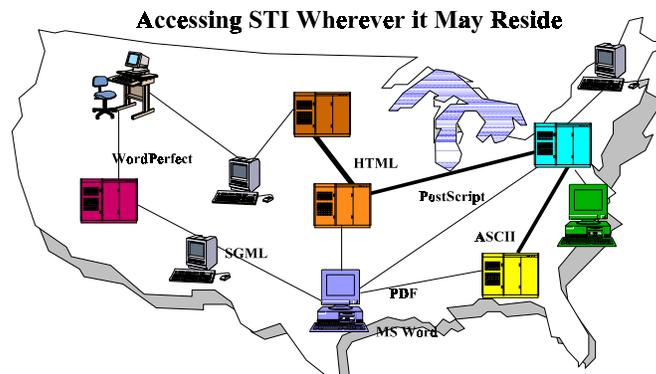


Figure 3 *The New Model*

retrieval across systems, greatly enhancing the availability and value of information in the research process.

Vision

Federated Collections provides an opportunity for those within the Department of Energy's scientific community to realize use of energy-related scientific and technical information (STI) through connected worldwide energy resources. This initiative, a component of the Department's EnergyFiles effort, will also explore how the researcher can access information collections, electronic journals and preprints, applied and engineering standards, database and document delivery services and regulatory, funding information and reference material, accessing STI where ever it may reside. In addition, state-of-the-art tools and technologies that facilitate scientific research and collaboration will be investigated and incorporated.

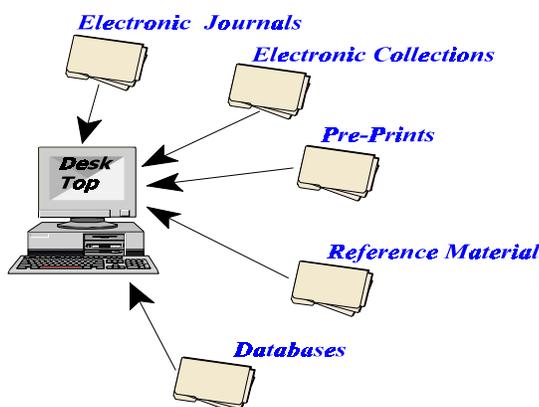


Figure 4 *An Objective of this Initiative is to bring Information to the Researcher's Desktop*

Architecture

The Internet leverages existing computer networks, extending and enhancing the ability of users to access information quickly and efficiently.

The Internet has become the universal mechanism for accessing and disseminating information. TCP/IP is the universal language of the Internet and is the basis for organizing and linking the vast amount of information available from the Internet. It is not a separate network but rather a set of information servers that can be accessed by client applications to access the data. This data can be text, pictures, sounds, video, or other multimedia format. A significant feature of the Web is its intuitive point and click interface, requiring minimal training, if any, to quickly become a competent user.

TCP and IP, two of the most popular communications protocols today, allow different kinds of computers using different operation systems to communicate with each other over a LAN or a WAN. The term TCP/IP is frequently used to refer to the collection of communication protocols that make up the full TCP/IP protocol suite.

The key facts that have contributed to TCP/IP becoming a universal protocol are the following:

- TCP/IP is based on open and well-published standards, thereby making it easy to write applications based on TCP/IP protocols for any type of operating system
- TCP/IP is not dependent on any specific operating system, and has been implemented for all types of computers, from PCs to mainframes.
- Many TCP/IP protocol implementations and applications are available in the public domain, reducing costs for implementation.
- TCP/IP works effectively and efficiently over a variety of LAN and WAN technologies.
- TCP/IP provides a common set of standard applications to provide services such as electronic mail, file transfer, and remote access to computers.

Because of the user friendly character and universality of the World Wide Web, it provides the optimal foundation for finding, accessing and using STI across the Department.

There are a number of approaches as to how data can be stored, accessed and shared across an architecture based on the Internet: for example, STI documents that reside in directories in the HTTP tree structure and are indexed on a routine schedule; local repositories of searchable indexes that contain pointers to full - text documents; repositories of STI documents that contain metadata that is searchable that are routinely indexed; a searchable repository of metadata that contains pointers to full-text residing in remote dispersed repositories.

For the Federated Collections Pilot, the team chose to establish a central index of searchable metadata with pointers to full text STI because we believe it provides the most efficient way of establishing the functionality needed to discover STI quickly and inexpensively over a distributed network of document repositories. It also entails the least cost, overhead and effort for organizations to participate. In one case, a participating organization's repository of STI was identified, and the OSTI technical staff quickly created a Python script to extract its Standard Generalized Markup Language (SGML) encoded metadata and add it to the central, searchable, index of metadata - all without any resources expended by the participating site.

Providing researchers and scholars with readily accessible information about the existence and location of materials held continues to be an important and valued feature of bibliographic databases; in contrast to the state of the Internet where the functionality of full text searching may be characterized as difficult, inexact and tiresome. And while the available information on the Internet is typically unstructured, increasing the difficulty of finding information, many information objects are simply not amenable to access via the Internet, such as repositories of paper, video, microfiche, microfilm and art where metadata may be the only means to determine the existence and location of these materials. As the enabling technology matures and becomes widely adopted for full text search across distributed repositories of information objects, bibliographic searching, which implies a standard metadata structure, we believe, remains the viable option of determining the existence and availability of information - and this, in conjunction with the availability of full text, provides an extremely powerful tool for our STI consumers.

Federated Collections will allow the DOE STI Community to:

- Publish once where many can access and read at the point of origin.
- Deliver information when the user wants it and in the form the user wants it.
- Facilitate sharing of information among information producers and information users
- Deliver quality information
- Manage scientific and technical information as a DOE-wide resource

Based upon the solicitation of user requirements and needs, this system should be based on the following requirements and characteristics.

- System response time should be within two seconds of entering queries for data
- System should be reliable
- System should be easy to use.
- Information should be easily shared among participants
- System should provide guidance, including online help
- System should be cross platform
- System should be scalable
- System should be reliable over WANs
- System should promote information integrity and security
- System should use commercial off the shelf or government off the shelf software where possible
- System should address heterogeneous document/data access
- Query Interface should be unified, consistent and seamless
- The system should be as non-invasive as possible, both to current and legacy collections.
- The system should ultimately be non-proprietary.
- Documents should be hosted and managed by the organization that created them

Federated Collections - Pieces to the Puzzle

- Shareable STI in DOE information systems
- The information repository and the access mechanisms used by participating systems

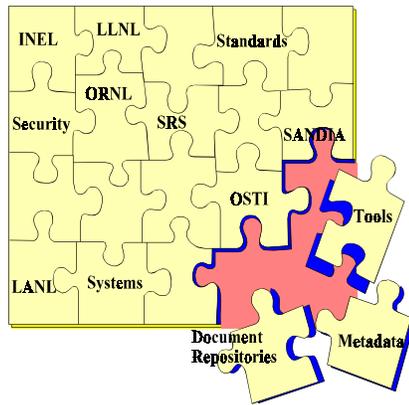


Figure 5 Pieces to the Federated Collections Puzzle

- Methods and tools, e.g., metadata, for accessing STI that are shared across the Department
- A repository system for storing this metadata and making it accessible at the point of origin
- Tools for creating, maintaining, and using the metadata
- Tools and procedures for controlling user and application access to common repositories of STI
- Standard tools and procedures for administering and maintaining shared repositories of STI
- Mechanisms for ensuring the security and integrity of the Department's STI.

What Have We Accomplished To Date?

The Los Alamos National Laboratory (LANL), Bechtel Hanford (BHI), SANDIA National

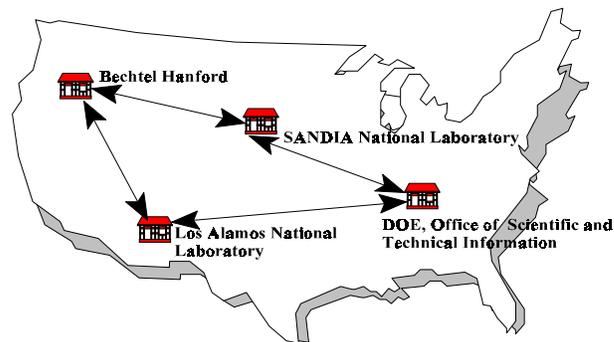


Figure 6 A DOE Office of Scientific and Technical Information and National Laboratory Collaboration

Laboratory (SNL) and OSTI have conducted a proof-of-concept pilot that takes the first step towards realizing a truly distributed capability to search via a single user interface among widely distributed repositories of information.

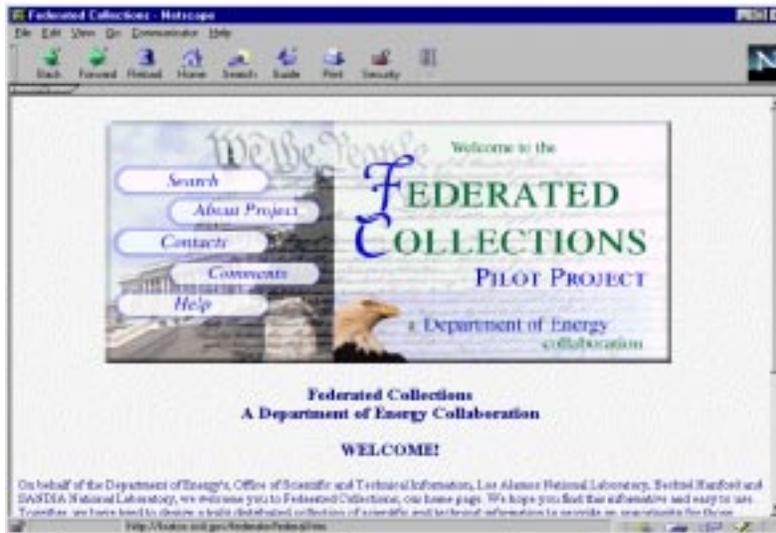


Figure 7 *The Federated Collections Home Page - <http://www.osti.gov/federate/federal.htm>*

Over a period of six months, each pilot participant provided a sub-set of scientific and technical documents to be used as a test information set. In addition to the pilot document test sets, LANL provided their knowledge and experience in search and retrieval technologies while SNL and BHI provided their significant expertise in technical information issues and management.



Figure 8 *Search Interface for the Federated Collections Pilot*

Meanwhile, the DOE, Office of Scientific and Technical Information contributed resources in the development of the user interface design and experience in deploying applications on the Internet.

The approach the Federated Collections Team agreed upon was to rapidly establish a working prototype and then continue to enhance it, evolving the prototype to reflect the requirements



Figure 9 Federated Collections Retrieval Results

and team vision. This distributed search and retrieval system is based upon Verity's Search 97 and Innovative Web Applications Explorer, using metadata extracted from the DOE, Office of Scientific and Technical Information's database of scientific and technical information bibliographic

```

<TITLE>Preliminary Hazard Classification for the 100-d
Site Remediation Project (Group 2)</TITLE>
<AUTH>D. K. Oestreich</AUTH>
<PUBDT>19961115</PUBDT>
<URL>http://www.bhi-
erc.com/library/bhi/bhi00913.pdf</URL>
<DOCTYPE>Report</DOCTYPE>
<AB>This document provides the preliminary hazard
classification for the 100-D Group 2 Sites Remediation
Project. This report provides project description
information in sufficient detail to support the assumptions
and conclusions as provided in this preliminary hazard
classification. The 100-D Group 2 Site Remediation
Project is targeted at excavation of contaminated soils
from nine waste sites in the 100-D Area and the
transportation of these wastes to the Environmental
Restoration Disposal Facility.</AB>
<DE>preliminary hazard classification; Hanford Site;
100-D; material-at-risk</DE>

```

Figure 10 Sample Metadata

Challenges

Federated Collections is the first, in a series of steps, toward building consensus on the technical and organizational requirements for future iterations of an increasingly sophisticated system of distributed and shared scientific and technical information repositories. The challenges are many and, not surprising, the least daunting of those challenges reside in the technical arena. With the right organizational support, processes, participants and technologies, Federated collections will grow to be a robust foundation for future expansion that will enable users within the Department to access STI collections where ever they may reside.

Challenges that Federated Collections should address in the near term are:

- Investigation, testing and implementation of a metadata standard
- Investigation, testing of information brokering technologies
- Full-text indexing
- Security - firewall issues, authentication, access and digital certificates/signatures
- What constitutes a record or true copy?
- PURL - or equivalent system to address the need to manage the problem of drifting URLs.
- Increased participation and commitment of stakeholders