Towards a Virtual Environment for Supporting Research Activities at the Smithsonian

Thornton Staples
Director of Research and Scientific Information Management
Office of the CIO, Smithsonian Institution

Abstract

The Smithsonian Institution supports research activities in all aspects of science and cultural heritage in both research institute and museum settings. This paper describes a conceptual framework and information architecture for the prototype repository-enabled virtual research environment that is under construction. The goal of the project is to support the researchers to get their information into a trusted repository as the first stage in the information lifecycle, then to be able to manage, analyze and disseminate the information in a linked-data world, retaining ownership and control until they are ready for it to pass to an institution to be curated for the long term. The prototype is based on Fedora and Islandora.

Introduction

The Smithsonian Institution was founded for the "increase and diffusion of knowledge" of all kinds. Though it was established as a research institution, researchers create and use collections of all kinds in their work. As it developed, preservation became the third principle of the mission. The Smithsonian now includes 19 museums, 9 research centers, 8 advanced study centers, 22 libraries, 2 archives and a zoo. Research activities, or the support thereof, goes on in all of these units.

The observation-based research that is associated with the sciences is a major part of the Institutions activities, with strengths in environmental sciences, biodiversity and astrophysics, as well as scientific research associated with the conservation of physical collections of all kinds. The museums, especially those associated with cultural heritage, host a wide array of research projects more involved with putting the collections in context, where many years of research can go into a museum exhibition or collection development. To this point, there has been no systematic effort to support and sustain the digital information that is created in both kinds of research activities.

In order to support all of these research activities, the Smithsonian has initiated a program to develop a repository-enabled virtual research environment. This effort is built upon the assumption that the scholarly and scientific record is rapidly evolving to become a formalized web of content, in which any node must be able to be linked to any other node, with formal, typed relationships. Rather than managing all of the digital information as an online version of a traditional library collection, the Smithsonian must be ready sustain its corner of a growing, organic network of information.

Creating and managing digital research information such that it can be sustained indefinitely, while made available to be easily used in elaborate and inventive ways, is no small task. The fidelity and authenticity of the information must be assured from the moment of creation and indefinitely throughout its useful life. A variety of policies must be enforced so that access is managed appropriately. The whole range of possible digital media must be supported, including but not limited to texts, images, video, audio, and tabular datasets.

Most importantly, the expression and organization of these research projects must be formalized and instantiated as digital information to be managed the same as the final research output. Research projects will almost always include multiple and various kinds of media files that each represent a facet of the project, but must be made sense of together to represent the project as a whole. Increasingly, any research project, or any of its parts, will be a source to be mined for data that can be use in new ways, for purposes not necessarily foreseen by its creator.

This kind of activity clearly requires some kind of institutional support standing behind it, taking responsibility for maintaining a corner of the world-wide network of information as durable digital content, while making it available to be used as linked data in all the necessary ways. Ideally, the hosting organization provides the infrastructure and support network that allows the researcher to build the project and see it through to its conclusion, while maintaining his or her control of the content and its use by others. After such a project has been concluded, it ownership must at some point be transferred to the host organization, or the project must be moved to another institutional home, to be sustained into the future.

The software tools, standards and practices that have resulted from digital repository management research, and from the continuing development of the Web, provide a foundation on which to begin to address this challenge. Though no standard exists for how to organize and manage the components of this scholarly network, the Flexible Extensible Digital Repository Architecture (Fedora) provides a conceptual framework that can address the basic needs.

The Fedora Commons implementation of this architecture provides a repository management platform that is well along an evolving path towards providing a foundation upon which to build. It provides the necessary components of information management, with one abstraction facing the storage of files, another facing the users of the content. The digital content is stored as a set of files that do not depend on the Fedora software to be understood. The necessary information needed to assure durability of each object is stored in files that are both machine and human-readable which can all be made available to software applications as flexible, policy-enforced, linked data.

Researchers will have to be directly involved.

The process of organizing digital information to be sustained indefinitely, while making it useful to both the researcher who is responsible for its creation and for others who may become interested in using it in new contexts, requires that a knowledgeable person make sense of it explicitly at the time of creation. Left to their own devices, researchers generally develop casual, implicit ways of organizing the information, using cryptic file and directory names, short-hand notes, and to a great extent their memories, to provide continuing access to the content. All of these methods quickly become untenable when the responsibility for management of the information changes, and completely hopeless in making the content discoverable and reusable by others for other purposes.

Any scheme to provide a way of organizing digital information to be sustained indefinitely and, if appropriate, to be reused widely, will require that the content's creator use a standardized approach and to explicitly provide the necessary description and classification of the content at the point of creation. Any approach that has a chance of succeeding will: 1) have to make it easy for the researchers to meet the requirements; 2) assure them that they will not lose control of the content until appropriate; and most of all, 3) give the researchers an immediate benefit from using the system that outweighs the effort that they need to invest.

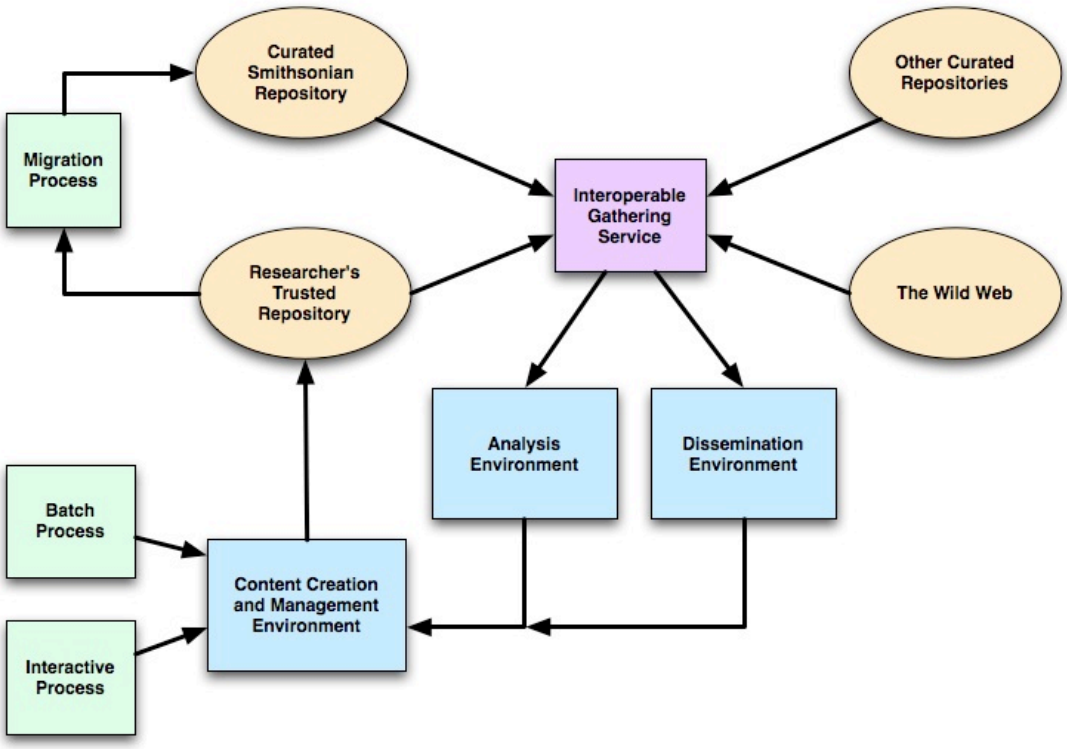

Figure 1.

Figure 1 shows a conceptual framework for how this could be accomplished. The basic assumption underlying this framework is that, over time, an environment of integrated software tools can be provided that meets all the researchers needs. The key to ensuring the sustainability and reusability of the information is to get it into a repository as early in the process as possible.

In the diagram the Researcher's Trusted Repository represents the initial state of the managed content, where the researcher who creates the content retains complete control of it and has the ability to set policies about how it can be used by others, or not. The Curated Smithsonian Repository represents the state of the information when it passes from the original researchers control to its necessary long-term institutional home. This may just be a logical separation, in which the data stays in the same repository and the ownership of the content moves from the researcher to the institution, or it may move

from one repository instance to another, maybe outside the boundaries of the original supporting institution.

The three "environments" represent virtual spaces where a user has access to software tools and services as sets of use-case specific interfaces and workflows. The Content Creation and Management environment handles the creation of new information objects in the Researchers Trusted Repository, both from interactive processes and batches of data prepared in some other way, as durable, trusted digital information. It also provides users with a way of updating their automatically versioned data, as well as the ability to manage the policies associated with it. The Analysis and Dissemination environments provide supported spaces where users can gather data from their own collections and from other repositories to process using an integrated set of software tools appropriate to the activity. In both cases digital content created in the process can smoothly flow back to the Researchers Trusted Repository to be added to their collections.

It is clear that the success of this effort will hinge on being able to provide a support environment to researchers that ensures immediate, highly useful, gratification that outweighs any extra effort needed. It is also clear that support of the complex array of research activities at the Smithsonian will not be possible unless a way is found to provide an integrated environment where software tools that work in standardized ways can be plugged together to ensure the smooth movement of data among all of the necessary states.

The Office of the CIO at the Smithsonian Institution has begun a pilot project using Islandora/Fedora to find the right starting point for creating a practical support environment that aims this conceptual model. The ability to manage a complex array of interrelated trusted, durable digital objects in a collaborative environment, under the control of fine-grained policies, comes with the package. The challenge is to create an architecture to organize the information and a "starter-set" of software tools in the three environments that exploits it to provide enough functionality to make life easier for a first set of early adopters. It should be emphasized here that this is not an effort to construct a complete system and roll it into place all at once. The repository and web-based architecture that will be developed can be grown more organically, starting with specific use cases, generalizing them, then adding more.

An Architecture for Managing Research Information

While the web is a great model for how information can be organized and shared, it is a terrible system for managing that information. The web is built by creating a backbone of structured text files (HTML) that provide some of the information that is rendered on the page, provide links to other information that can be rendered on that page (like images), then provide the context for links to other web pages. These pages and related content are then strung together in a network, based on a simple set of rules implicit in how the server interprets the HTML text. These rules are mostly related to rendering the information on the page, not about abstractly organizing the data for any possible use. What is needed is an architecture that exploits the same information object and relationship orientation, formally instantiating those objects and relationships so that they are both durable for the long-term and so that they can be exploited in any number of ways.

Fedora provides elaborate ways to ensure the fidelity and authenticity of the data, manage both the content and its related metadata together, and to formally manage all of the relationships among the "information objects" that are part of the research project. Each of these objects can assert a formal, typed relationship to any other object. In it's simplest form, this system could be used to create a formally structured web of related objects that describes the research project and all of the digital content that relates to it. Storing the images, audio, video and datasets as objects in the repository is routine at this point. The challenge is providing the backbone of objects that creates the project context for all that content, equivalent to web pages but acting more like a flexible, organic database.

In the set of objects for a research project, the "home page" would be a conceptual object representing the research project as a whole, a structured text that is essentially a database of the top-level information about the project. It could include a formal title, short descriptive texts that explain different things about it, a variety of metadata fields that classify it, describe people and institutions associated with the project, etc. The research project object could then assert relationships to media

objects that contain such things as sets of notes, emails associated with the project, texts (such as grant proposals, essays, articles that result from the research, etc.), and digital information created for the project (such as images, audios, video and tabular datasets), as shown in figure 2.

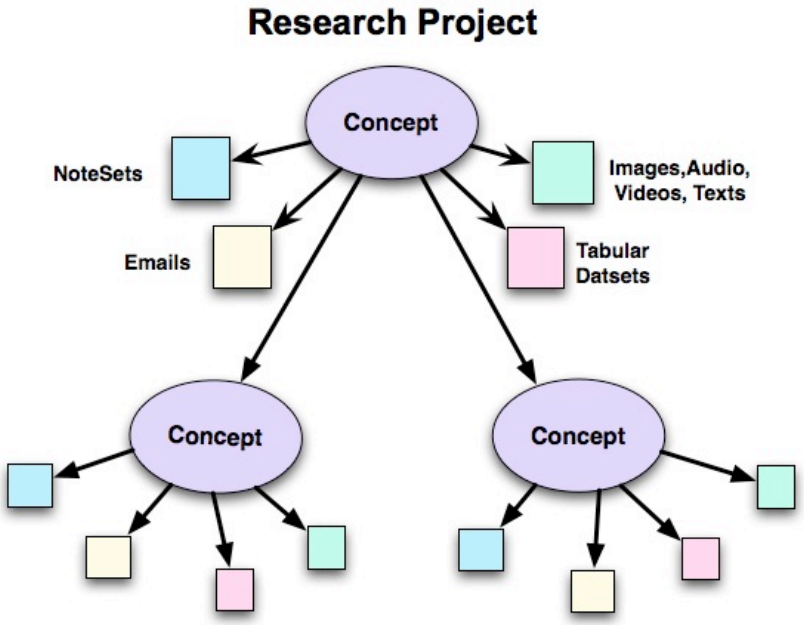

Figure 2.

Like an HTML page for the web, these conceptual objects would contain an XML file that can be rendered to create a web page, as well as to be indexed to make the research project discoverable as appropriate. In its simplest form this could easily be accomplished using an appropriate descriptive metadata schema. Where information needed to be exploited to order the child objects in a specific way, a more complex schema, such as the Metadata Encoding and Transmission Standard (METS) or the Encoded Archival Description (EAD) could be used, or a new schema could be developed.

The conceptual object would use the standard functionality of Fedora to store each relationship that it asserts to its child objects as RDF triples in its RELS-EXT datastream, ensuring that the relationships would be indexed in the repository's Resource Index, and that the set of all of its relationships is equally treated with its content and metadata. This would mean that the research project object's identifier could be used to find all of its children. The XML data in the object could be rendered as the homepage of the project, and links to the children could be provided. By exploiting this functionality, a default web interface can be provided that allows the researcher to browse through all of the data for the project.

Other conceptual objects could also be children of the project object, allowing for a more complex structure to be built for the project. Each conceptual object could have all of the same kinds of media objects as its immediate children.

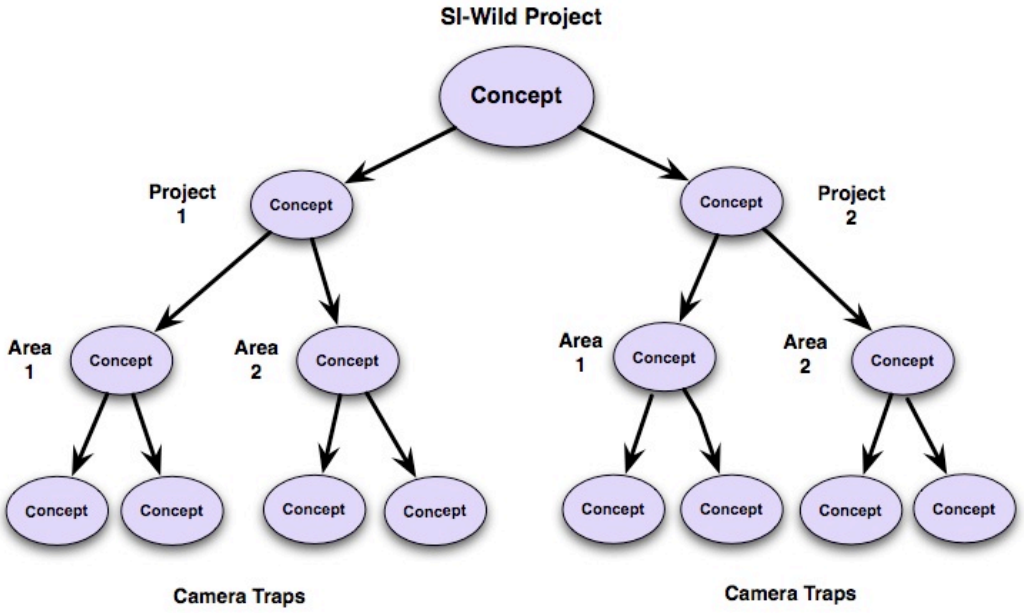

Figure 3.

Figure 3 shows an example of how this works for one of the projects in the first prototype. The Smithsonian Wild[1] project is an international project among a variety of organizations for which the data is hosted at the Smithsonian. The project studies the distribution of animal species by collecting candid photographs from motion-sensitive camera traps day and night. The researchers capture metadata about each of the images that describe the animals in the pictures, time, place, etc., which is then analyzed to gather statistics about the populations as a whole, as well as to study individual animals over time.

The project is organized by a set of conceptual objects, with a root object that represents the project as a whole. Within the main SI-Wild project there are individual project groups, working in different regions, each of which has different areas that they are studying. Each area has a number of camera traps, each of which is the parent of a set of images that grows over time. Each of the conceptual objects contains metadata appropriate to that level. For example, the camera trap objects have data about the location of the trap, who is responsible for it, and configuration information about the camera itself. The ability for Fedora objects to be automatically versioned is very useful here, as the date of the images related to a particular camera can be easily reconciled with the configuration information from that date. Each node could have other kinds of objects associated with it as well, such as text objects that are grant proposals and papers, or tabular datasets that contain statistics derived from the photographs which are used in publishing, etc.

Conceptual objects that describe museum objects in the context of the research project are also important to the Smithsonian. These objects could get their initial data from the appropriate collection management system and source their images in from the Digital Asset Management System (DAMS), but they would provide an enduring version of the physical object's description relative to the project. It is critical that the description of the object initially provide the information that is current in the collection management system, but then becomes a record of the description of that object relative to the project. Researchers often change the understanding of the object, but even if they do not, the description in the collections system can change after the fact. All of these conceptual objects would lend themselves to being the source of different web pages for different purposes.

[1] SI-Wild http://siwild.si.edu/

Figure 4 shows how the "Posters American Style" exhibition[2] done at American Art in 1998 could have been managed this way.

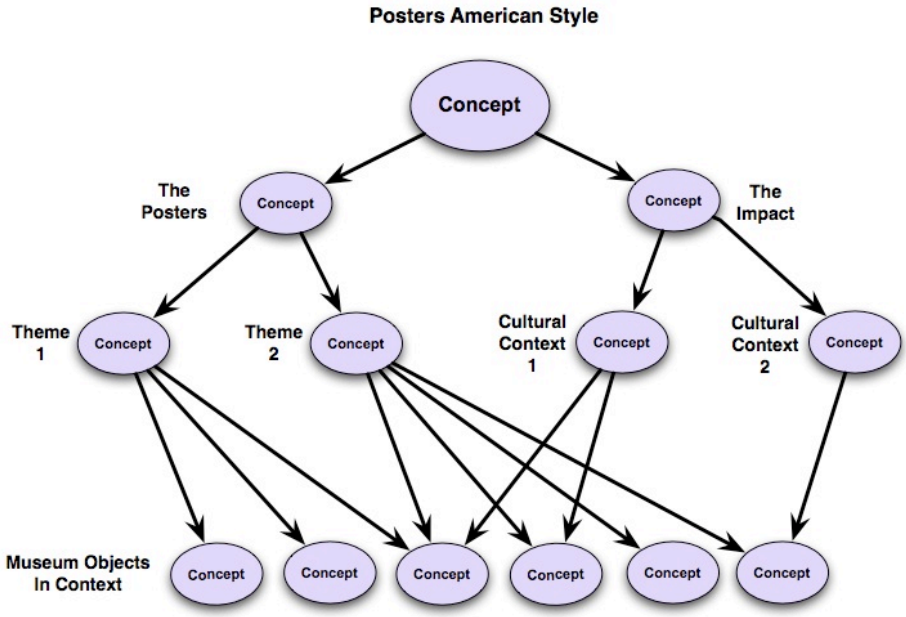

Figure 4.

The exhibition is organized into two main areas, "The Posters" which groups the collection of posters to be exhibited into four thematic areas, and "The Impact" which highlights posters in terms of cultural events or movements. There are essays for the exhibition as a whole, for each of the two major divisions, and for each subdivision thereof. Each thematic area then relates to digital surrogates for a group of posters, some relating to more than one thematic area. Each of the impact areas refers to one of the posters. There is also a section of the exhibition not reflected in figure 4 that describes the process of making posters, which has images and animations that are related.

The Carrot: Providing Immediate Benefit to the Researcher

The most immediate benefit to the researchers by taking this approach is to have their data organized as a navigable network of interrelated objects. Providing an interface that allows them to navigate up and down the tree of conceptual object nodes provides the basis. At each node the user could see both the conceptual description from that object rendered on the page, as well as lists of links to the media objects that are its direct children, which provides direct access to both the metadata and content of each of them.

The network of objects will exist within a policy-controlled framework that allows the owner of the objects to set and maintain the policies. This can be a single researcher with a relatively simple project who can make their data available (or not) to colleagues or the public, if desired. A research group with many collaborators, like the SI-Wild example above, could distribute the management of policies for subsets of the project to different researchers. The researchers retain these rights until the ownership of the project is passed to the institution.

The workflows and interfaces that allow the researcher to create the conceptual objects, and to ingest and describe the media objects as painlessly as possible are very important. Islandora provides a good bit of what is needed, including the ability to create and edit XML data from datastreams in Fedora objects for a variety of metadata standards, and some of the workflows and interfaces that are needed to

[2] Posters American Style http://americanart.si.edu/exhibitions/online/posters/mainmenu.html

ingest and describe sets of media objects. The basic functionality for creating and relating conceptual objects to create the backbone of the network described above is there. Clearly, the incentive to users is to be able to use their data, along with data that they gather from other sources, in the analysis and dissemination environments, to gain an immediate advantage for the work that they have to do.

A whole variety of repository processes and tools could be provided that would exploit these organized networks of research data. There are many more possibilities that could exploit this structure, but here are a few:

- A researcher who provides some descriptive metadata and descriptions for the columns in an Excel spreadsheet which they upload could immediately download a version that was ready to use in statistical programs such as SAS, SPSS or R, or as an attribute table in a GIS model.
- A tool that could selectively harvest a research project to provide a direct connection from the researcher to the museum departments which create the gallery exhibition and the one that goes online.
- A drag-and-drop environment that allows researchers to gather datasets from the repository and drop them in to a process where they could create a new aggregate dataset for their work or to bring them into a model.

A great deal of work has been done in the virtual research environment community that could be applied to the analysis and dissemination environments in this model. The challenge here will be to find ways to smoothly extract data from the repository in the form needed to use it as input to software tools used in those environments.

Conclusion

This project is clearly very ambitious in its long-term goals, but is designed in such a way that it can initially be grown among a small number of researchers, then expanded. The prototype now in hand provides enough of the described functionality to begin working with an intrepid group of researchers who are early adopters, to be able to work towards developing the level of functionality that will ensure wider adoption.

It is clear that an approach based on the metaphor of a library, a perfect collection with its perfect search interface, is inadequate for the task. It is just as clear, that slapping a layer of RDF-based linked data on top of the existing World-Wide Web is not the answer either. This project aims to be somewhere between, providing trusted, durable digital information in the initial stage of the research data lifecycle, and growing it as part of a web-based scholarly record that is truly interoperable.

Researchers will have to have to become directly involved in describing and organizing the information that results from their activities. Getting researchers to do these kinds of things has often been compared to herding cats. It has also been said that the only way to herd cats is to tilt the floor. Funding agencies around the world have begun to tilt the floor with their requirement for data management plans for research grants. Thirty years ago most researchers did not type their own papers; technologic developments that gave them complete control and ensured their convenience changed that paradigm. It remains to be seen whether or not they can be enticed into creating their own metadata and organizing their own content.