Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computers and Electronics in Agriculture 47 (2005) 221–242
Computers and electronics in agriculture
www.elsevier.com/locate/compag

# The open research system: a web-based metadata and data repository for collaborative research

Charles M. Schweik [a, *], Alexander Stepanov [b], J. Morgan Grove [c]

[a] *Department of Natural Resources Conservation and the Center for Public Policy and Administration, University of Massachusetts, 217 Holdsworth Hall, Amherst, MA 01003-9285, USA*
[b] *Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst, MA 01003-9285, USA*
[c] *USDA Forest Service, Northeastern Research Station, 705 Spear Street, South Burlington, VT 05403, USA*

**Abstract**

Beginning in 1999, a web-based metadata and data repository we call the "open research system" (ORS) was designed and built to assist geographically distributed scientific research teams. The purpose of this innovation was to promote the open sharing of data within and across organizational lines and across geographic distances. As the use of the system continued, end users and group administrators requested the development of a second, Intranet-based system with similar functionality. After three years of operation, a survey was conducted of users of the system to understand why some users and research groups appeared to utilize the two systems more than others. From this research we found that some barriers to use include: (1) mismatch of system functionality to user or group needs; (2) willingness to share with an internal group by Intranet but not with the world by Internet; and (3) resistance to entering metadata because of workplace habits. This experience has also taught us that with respect to web-based metadata and data repositories there is a difference between long-term and short-term research projects in their need to establish good metadata and data storage procedures. Moreover, some time is required for researchers to change from short-term to long-term project thinking. It is also important for organizations or managers of such research groups to reflect on established incentives and penalties that either encourage or discourage appropriate use of metadata in filing procedures. We conclude with a discussion of possible improvements that will be made to the system in the coming years, with an emphasis on the emerging phenomenon of "open content" (OC) collaboration that is being modeled after Internet-based collaboration in "open source" (OS) programming. This development will require online systems like ORS, and the OS/OC approach has

* Corresponding author. Tel.: +1 413 323 7682.
*E-mail address:* cschweik@pubpol.umass.edu (C.M. Schweik).

0168-1699/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2004.12.006

the potential to evolve into no less than a new paradigm for how cross-organizational (in fact, global) scientific research collaborations are undertaken in the future.
© 2005 Elsevier B.V. All rights reserved.

*Keywords:* Group collaboration; Metadata management; Online database; Open source; Open content licensing

## 1. Introduction

Geographically distributed scientific research and collaboration, as we know it, began in the 16th century with the advances of the printing press and the postal system (Ziman, 1969; Johns, 2001; Lucky, 2000). Peer-review emerged as an important method to improve the quality of scientific information (Burnham, 1990; Kronick, 1990). This system has been the dominant form of scientific collaboration ever since, but it is a formal system for sharing research results, rather than being a true collaboration across organizational lines.

In the 1960s and 1970s, early advances in computing technologies and networking (in particular, the development of Internet foundations such as Arpanet) created new opportunities for scientists to collaborate more easily from remote settings. For example, in the early to mid-1970s, cross-university collaborations were established to work on energy-related research (Estrin, 2000). But these early days of Internet-based collaboration required scientists to have technical knowledge of how to communicate. Over the last decade, advances in email (and attachments) have made Internet-based collaboration more viable for scientists outside of technical fields like computer science (Lucky, 2000). Some studies on email use report increased productivity and enhanced communication among workers (Finholt et al., 1990; Walsh and Bayma, 1996; Cohen, 1996). Other studies highlight problems with email-based versus face-to-face or phone communication such as higher levels of misunderstandings among participants (Sproull and Kiesler, 1991) and, also, work disruptions (Ancona and Caldwell, 1990).

These same decades also exhibited substantial growth in the availability of network-accessible databases and digital libraries. Online research databases moved from mainframe-based bibliographic databases with search queries of the 1960s and 1970s to the distributed databases with full text search capabilities that constitute the web environment of today. Many government agencies now maintain large-scale data servers and metadata "warehouse" systems (Sen, 2004) in an effort to make their data more publicly available over the world-wide web (see, for example Kerschberg et al., 1996 which describes a U.S. National Aeronautics and Space Administration system, or Gillman et al., 1996 on a U.S. Census Bureau system). And recent projects like the National Biological Information Infrastructure or the Government Information Locator Service provide examples of large-scale government-driven projects which make access to data easier for various stakeholders (Sepic and Kase, 2002; Moen, 2001). Some particularly advanced metadata and/or data sharing systems include the U.S. Geological Survey (USGS) EarthExplorer system (http://edcsns17.cr.usgs.gov/EarthExplorer/), U.S. state agency repositories such as Massachusetts' MassGIS system (http://www.state.ma.us/mgis/), and meta-search facilities such as the U.S. National Geospatial Data Clearinghouse (http://clearinghouse1.fgdc.gov/).

However, particularly from around 1995 onward, the combined capabilities of email, web pages and web-connected databases have led to the emergence of research collaborations among groups of researchers who no longer work in a common geographic location (or even a common organization), and work together as a "virtual team" (Lipnack and Stamps, 1997). One group with whom we are affiliated—the Baltimore Ecosystem Study (BES), a Long-Term Ecological Research (LTER) organization funded by the U.S. National Science Foundation (NSF)—provides an example of such a geographically dispersed, virtual team. The more than 50 affiliated scientists participating in BES research projects are physically based with organizations scattered across the eastern United States, including a variety of universities, government agencies, and other non-profit organizations. This is not a group organized by a single government agency, but rather is an interdisciplinary research group with members who are either formally (through research funding) or more loosely connected to the research program.

The ability to access expertise without being constrained by geographic proximity certainly is a benefit of this Internet-based collaboration trend, but it also raises some challenges. In particular, the distributed nature of virtual teams, coupled with the continued gains made in desktop computing, leads to a serious challenge in information and data management. As early as 1993, B.W. Hesse and colleagues called attention to this problem when they wrote: ". . . in one large project, despite international data centers, data coordinators, and approved data formats, 'scientists have their own data on personal computers in their favorite spreadsheet database"' (Hesse et al., 1993, p. 92). This statement articulates the challenge faced not only by BES, but by many such research groups: How can a team manage and share their information resources among members in a decentralized and geographically scattered virtual organization? And how can this team effectively utilize the Internet to share research findings with interested parties outside of their virtual team? This second question is an important one for government-funded research projects (such as NSF-funded programs) and government agencies, who are often mandated to share their findings with the general public as a requirement for their funding. The project described in this paper addresses these questions.

In 2000, our goal was to find an existing web-based metadata and data repository that could be adapted to improve the long-distance collaboration and data sharing among members of affiliated research groups and to encourage data sharing across organizational lines. Related to this last point—the sharing across organizational lines—we were intrigued by the open source phenomenon that was emerging in the field of computer science and was producing software such as the Linux operating system and the Apache Web Server. In our view, this new form of collaboration grounded upon innovative intellectual property licensing such as the GNU General Public License and Internet-based forms of collaborative infrastructure could provide a model for scientific research collaboration (Schweik and Grove, 2000; Schweik and Semenov, 2003). Collaborative development of computer programs is in many respects similar to collaborative development of scientific research. All are some form of intellectual property. Consequently, to the degree possible, we wanted to develop a web-based metadata and data repository that followed some of the principles of programming projects (note that there is more on this particular issue in the conclusion).

After searching and finding nothing readily available for the needs of the project, the project team developed its own system. This paper describes the design, development and

maintenance of the open research system or "ORS". Currently, ORS refers to two web-based metadata and data repositories designed to facilitate collaborative research over the Internet: http://www.Open-Research.org and http://www.Orsprivate.org.

Three parts of the paper follow. First, we describe the original intentions for the ORS system (http://www.Open-Research.org) and its technical design. We also explain why a separate Intranet version of the system (http://www.Orsprivate.org) emerged, and we detail its design and capabilities. Second, we reflect upon several years of operation and discuss why use of these systems has been somewhat limited. We report some of the findings from a survey of user group "administrators" and a sample of users. Third, we conclude the paper with reflections and insights from this effort that will be of broad interest to others, particularly readers interested in establishing their own Internet-based data management and research collaboration systems. We also discuss some of the enhancements to ORS that are currently underway, and describe in more detail the emerging paradigm of "open source/content" collaboration, for this an important direction we think organizations will take over the next decade.

## 2. The open research system (ORS)

### *2.1. The design of the initial open research system: http://www.Open-Research.org*

The initial concept of ORS emerged in 1999 as part of a broader effort by researchers at the USDA Forest Service's Northeastern Research Station to develop an integrated and collaborative research program studying social and economic components related to forests in the northeastern United States. In the early stage of this project, the team identified a critical obstacle: the need for a multi-scale, regional database and website to facilitate integrated and collaborative research among researchers in diverse disciplines and working out of various research offices distributed across the northeastern United States.

Behind this idea was a practical public administration problem—what we refer to as the "file cabinet problem"—which many organizations face. This is the problem of individual office file cabinets which house valuable research products and data, but which, because the filing system is personal and haphazard, become unusable when employees retire, die, or leave the organization. With this consideration in mind, the initial requirements of ORS emerged. The system should provide affiliated organizations with: (1) a mechanism to store metadata about datasets and papers stored in an organization's files; (2) a web-based searchable database of the metadata; and optionally (3) a way to upload the associated datasets and papers themselves that could then be directly available to people outside the affiliated organizations. These non-affiliated ORS users are broadly defined. Our intention was to make these metadata and datasets available to anyone who might have an interest in the data. For example, these users could be people in other government agencies or non-profit groups with an interest in the subject or geographic region.

At the time, no available web-based metadata repositories met our requirements. Consequently, we developed our own. At the same time, we were aware of the open source programming phenomenon occurring in the field of computer science, and we realized that the principles behind this effort could be applied to other collaborative endeavors beyond

computer science and leading to more innovative production of new knowledge (Schweik and Grove, 2000; Schweik and Semenov, 2003). Our thought was that by following open source principles and providing a web-based collaboration platform, it might be feasible to create a system that promotes the sharing of data, publications and other information among organizations interested in some aspect of forests in the northeastern United States.

There are many possible functions that might be provided in an Internet-based collaboration system (see, for example, http://www.fullcirc.com/community/communityfacilitation.htm; or http://www.bowlingtogether.net/). Table 1 provides a list of current functions Open-Research.org provides and compares this with several other web-based collaboration platforms. The table is based on a review of online literature about systems of which we were aware during the spring of 2003. We first categorize functions by "synchronous" and "asynchronous" collaboration. Synchronous functions allow people to collaborate or communicate online in some immediate fashion, such as in the context of chat rooms, online conferencing facilities and group "white boards". The asynchronous category is broken down further, into "coordination/collaboration", "data management", "feedback", and "settings/administration". In Table 1 we organize specific collaborative functions within each of these subcategories. Because of the importance and need for an Internet-based system for file and data sharing as part of our initial goals, our Open-Research.org system development efforts focused on the asynchronous data management functions marked with an "x" in Table 1.

The technical design of the Open-Research.org system is presented in Fig. 1. There are two world-wide web entry points: (1) the end-user web interface (Open-Research.org's functions) identified at the top of Fig. 1 and (2) the completely separate system administration web interface shown at the bottom half of Fig. 1.

On the end-user side (grey areas at the top of Fig. 1), three primary functions are provided: (1) user registration; (2) submit information (metadata with an optional upload data function); and (3) search (metadata). All web pages and functions are programmed using ColdFusion™ and feed a relational database for metadata storage. The registration function allows a new user to create a system account, which then can be used to submit new data.

The "submit information" option (Fig. 1) allows users to enter metadata related to four data types: geographic, non-spatial, publications, and web-reviews. Geographic data includes any kind of data that are spatially explicit; presently much of the data references geographic information systems (GIS) layers. The metadata fields included in this form are based upon minimum requirements of the Federal Geographic Data Commission standards (http://www.fgdc.gov/) and the design follows some of the "MetaLite" software structure provided by the USGS and the United Nations Environment Programme (http://edcnts11.cr.usgs.gov/metalite/). Over time, slight modifications to this structure have been made as a result of dialogs with ORS users. The "submit non-spatial" metadata option (Fig. 1) provides users with a way to enter metadata for non-geographically explicit data. Examples of this kind of data include spreadsheets, presentation files, graphics (e.g., .jpg or .gif files), ASCII text files, or any kind of data or working "content" someone might want to share with others. The "submit publication" metadata option (Fig. 1) provides a mechanism to submit bibliographic references to offline or online peer-reviewed and published papers. Finally, the "web-reviews" option (Fig. 1) provides a method for others to submit website addresses of interest to the broader group along with some descriptive and evaluation meta-

Table 1
Functionality of open research system and other web-based collaboration systems (as of Spring 2003)[a]

| Functions provided | Open-Research.org | Orsprivate.org | groups.yahoo.com (Yahoo) | www.AOL.com | www.MSN.com | www.LiveJournal.com | www.Share360.com | www.Source Forge.net | E-room |
|---|---|---|---|---|---|---|---|---|---|
| Asynchronous | | | | | | | | | |
| Coordination/collaboration | | | | | | | | | |
| Project management | | | | | | | | | |
| Calendar | | | x | x | | | x | | x |
| Memos/notes | | | x | | | x | x | x | |
| Listserver | | | x | | | | | x | |
| Threaded discussion lists/bulletin boards | | x | | | | x | | x | x |
| Event scheduling | | | x | | | | | | |
| Data management | | | | | | | | | |
| Data repository | x | x | | | | | | x | x |
| Information sharing within group | | x | | | | | | x | x |
| Information sharing with general public | x | x | | | | | | x | |
| Search data/metadata | x | x | | | | | | x | |
| Submit/upload data or changes | x | x | | | | | | x | |
| Version control | | | | | | | | x | |
| Export data | x | x | | | | | | | |
| Feedback | | | | | | | | | |
| Online surveys | | | x | x | x | | | x | |
| Feedback/web-email | x | x | | | | | | x | |
| Settings/administration | | | | | | | | | |
| Security access based on user/group privileges | x | x | x | x | x | x | x | x | x |
| Synchronous | | | | | | | | | |
| Online collaboration | | | | | | | | | |
| Conferencing/chat room | | | x | x | x | | x | | x |
| Instant messaging | | | x | x | x | | x | | x |
| White boards/one-to-many presentations | | | | | | | x | | x |

[a] *Note*: This table is not meant to be a comprehensive list. It is only a sample of some important and popular group collaboration web sites. Based on an online review during the spring of 2003.

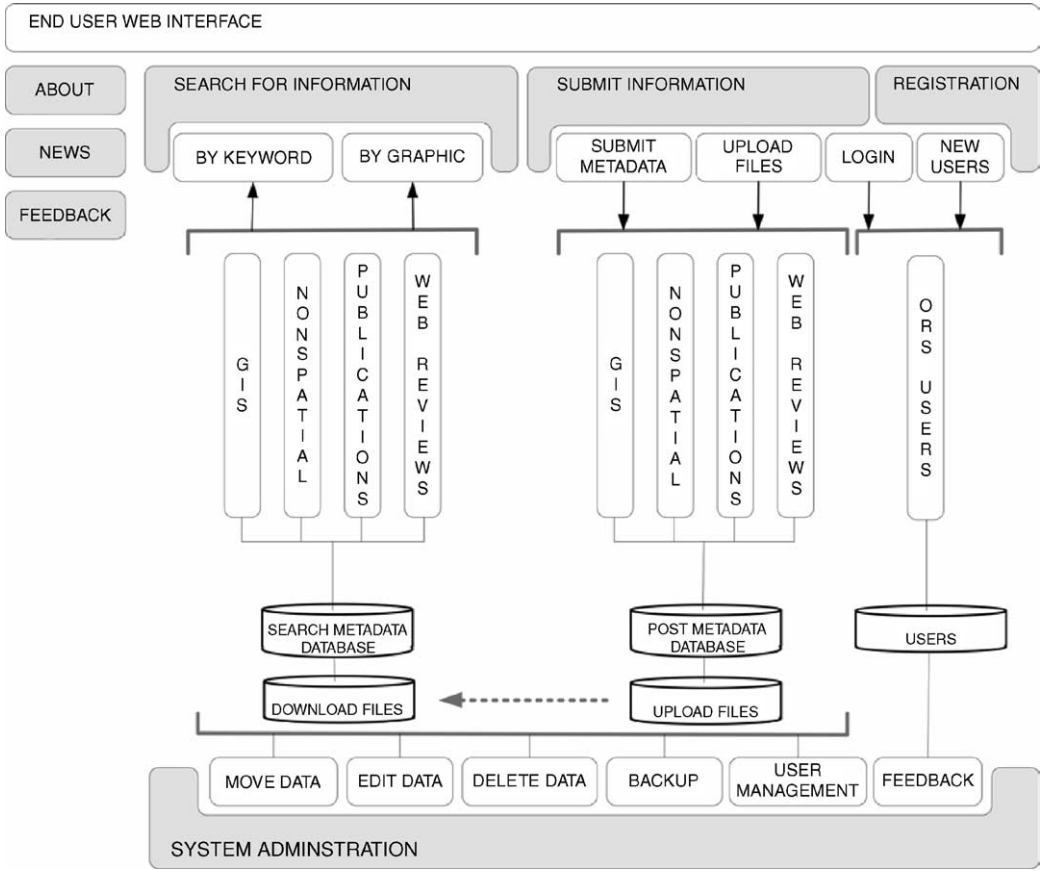


Fig. 1. The technical design of the Open-Research.org system. There are two means of entry: (1) the end user interface and (2) the system administration interface. Main functions for users are searching and submitting of information (metadata). Submitted data are stored in the post-database and are reviewed and moved by the administrator to the search database.

data about the site. In each of these metadata functions, the system provides the option for the user to upload up to five files to be stored on the server, the option to provide a URL to another server, or the option to provide a description of where the data are located if they are not online anywhere, and who to contact for more information. If the actual data are on another Internet server, a hyperlink in the metadata allows others using the search function (described next) to navigate directly to this location after reading the metadata. If the user decides to make data available to the public through the upload function, the data are scanned for viruses and then compressed.

While the submit functions in Open-Research.org require a user to register with the system, the functions within the "search for information" option do not (Fig. 1). ORS provides two search mechanisms, a standard keyword search facility and a "graphical search" function that is based upon a theoretical framework guiding the research of a major user group of the system—the BES LTER group. Users can search on all the different types of metadata in the system—GIS, non-spatial, publications, web-reviews—or can restrict the search to only categories of interest. They can sort results by author, title or keyword, and they can search data from all organizations using ORS or can restrict it to search for data only from one group or organization. Fig. 2 provides an example of the Open-Research.org end user web interface (described in Fig. 1) based on the search results for a GIS query.

A major issue behind our initial design of Open-Research.org was the question of how to make a system that was "open" for the sharing and submission of metadata and datasets themselves by interested parties within affiliated research groups and outside of the formal group (following some of the general principles emerging out of the open source programming movement), yet providing adequate protection against the submission of spurious metadata. Our solution was to divide the metadata database and associated upload files into two sub-databases: a "post metadata" database associated with the submit functions, and a separate "search metadata" database associated with the search functions (Fig. 1). When users post new metadata records through the online web forms, data are stored in the "post" database. The responsibility then lies with the Open-Research.org administrator, who has special system authority and access through the system administrator web interface (lower half of Fig. 1) to periodically review posted records and make decisions about whether to move them to the search database, to make further edits to the metadata before moving them, or to delete records that are not considered appropriate for the system. In essence, the administrator acts as a group moderator, responsible for checking the accuracy of all postings. This separation of post and search databases and the administrator role is, in our view, a critical design component that future online database and collaboration system development efforts should follow if they wish to adopt "open" principles while at the same time protect end users from accessing errant data on the search side of the system.

### *2.2. The emergence of an Intranet system: Orsprivate.org*

After approximately 18 months of Open-Research.org operation, it became clear that many of the existing and potential users of ORS wanted an Intranet version of Open-Research.org; one that would assist them in their collaborations with other remote researchers, but only with each other and not the general public. At the request of the user

Fig. 2. The search results page in Open-Research.org for GIS metadata. *Note*: This is an example of one of the pages of the end user interface shown in the top of Fig. 1. Clickable icons allow the user to: (1) get access to the uploaded files (diskette icon available or "off"); (2) to view the complete metadata (the paper icon); or (3) to view an abstract describing the dataset (the post-it note icon).

community, a second system based on the original ORS platform was developed, called "Orsprivate.org", which provides groups with this Intranet functionality.

Orsprivate.org offers the same features as the original Open-Research.org (e.g., submit, search, administrator functions), with the exception that it is designed around the concept of private research groups. The system architecture for Orsprivate.org is shown in Fig. 3. It is fairly easy to establish a new group in the system. All it requires is that the new group name be added to the "groups" portion of the database, and a specific URL and tailored homepage constructed so that the new group entry point can seamlessly interface with the websites of research groups already in place.

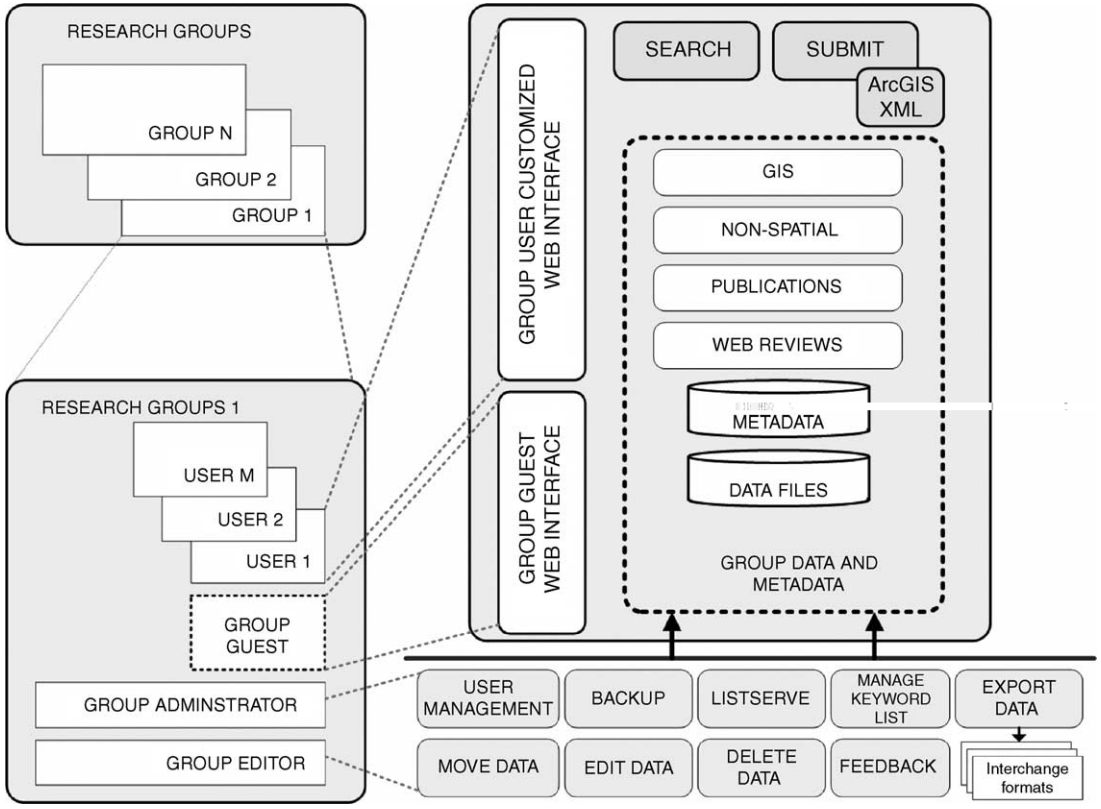


Fig. 3. The architecture of the Orsprivate.org system. Research groups are initially defined in the system by the ORS system developers (top left). A "group administrator" userid is created for each research group. Group administrators (bottom of figure) can create or delete group users, move submitted metadata from the group's submit to the search database, can edit metadata submissions, manage group keywords (used in metadata submission) and export metadata to other formats such as text files. General group members or users can search and submit the online database for metadata related to their particular project or group. Orsprivate.org also allows guest access through a guest interface for non-group members to search for data that is marked as "for public consumption".