Available online at www.sciencedirect.com
SCIENCE DIRECT®

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).

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.

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.

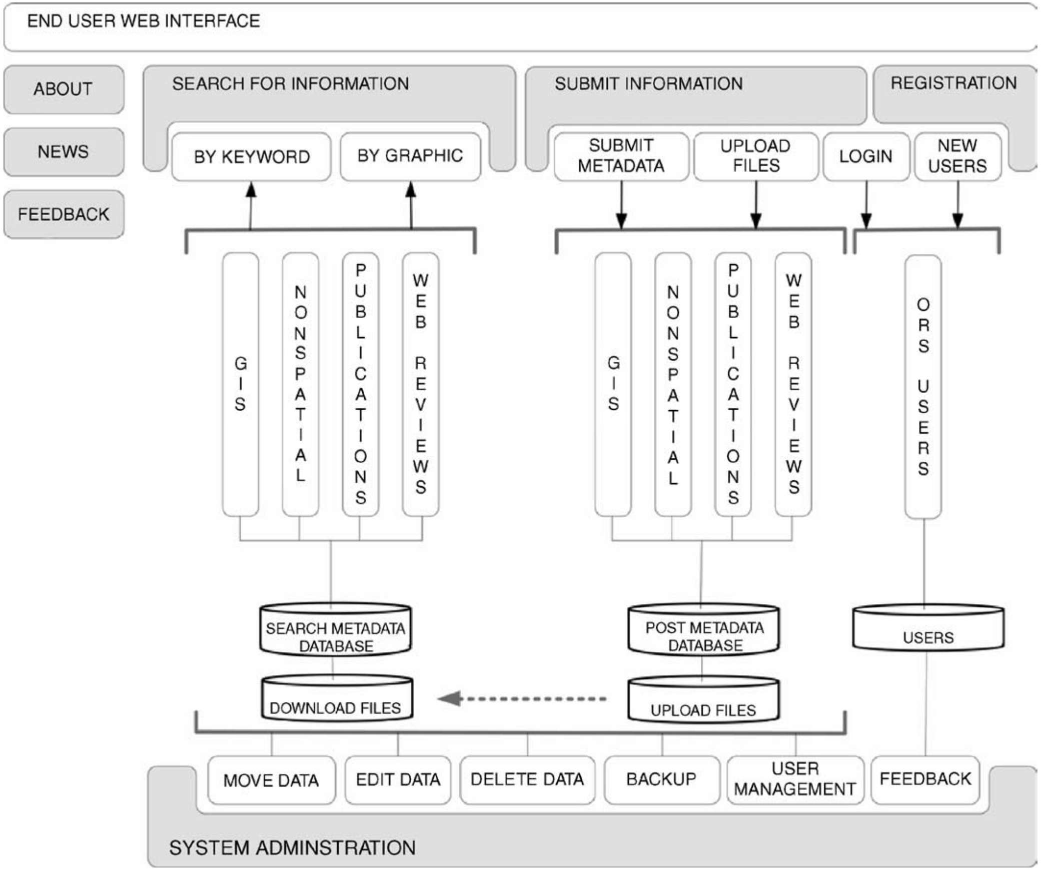
On the end-user side (grey areas at the top of Fig. 1), three primary functions are provided: (1) user registration; (2) submit information (metadata with an optional upload data function); and (3) search (metadata). All web pages and functions are programmed using ColdFusion^{TM} and feed a relational 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.

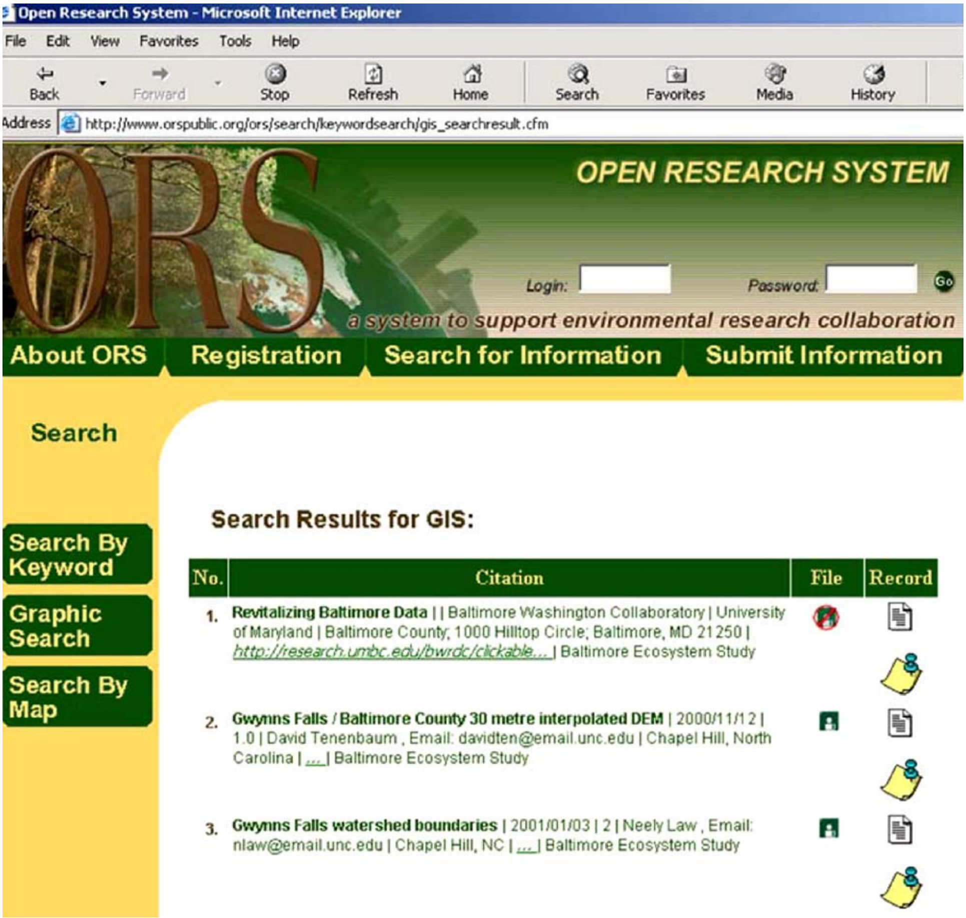
data about the site. In each of these metadata functions, the system provides the option for the user to upload up to five files to be stored on the server, the option to provide a URL to another server, or the option to provide a description of where the data are located if they are not online anywhere, 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


Open Research System - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History
Address http://www.orspublic.org/ors/search/keywordsearch/gis_searchresult.cfm
ORS
OPEN RESEARCH SYSTEM
Login: Password: Go
a system to support environmental research collaboration
About ORS Registration Search for Information Submit Information
Search
Search By Keyword
Graphic Search
Search By Map

Search Results for GIS:

No.	Citation	File	Record
1.	**Revitalizing Baltimore Data** | | Baltimore Washington Collaboratory | University of Maryland | Baltimore County; 1000 Hilltop Circle; Baltimore, MD 21250 | *http://research.umbc.edu/bwrdc/clickable...* | Baltimore Ecosystem Study	⊘	📄 📝
2.	**Gwynns Falls / Baltimore County 30 metre interpolated DEM** | 2000/11/12 | 1.0 | David Tenenbaum , Email: davidten@email.unc.edu | Chapel Hill, North Carolina | ... | Baltimore Ecosystem Study	💾	📄 📝
3.	**Gwynns Falls watershed boundaries** | 2001/01/03 | 2 | Neely Law , Email: nlaw@email.unc.edu | Chapel Hill, NC | ... | Baltimore Ecosystem Study	💾	📄 📝

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.

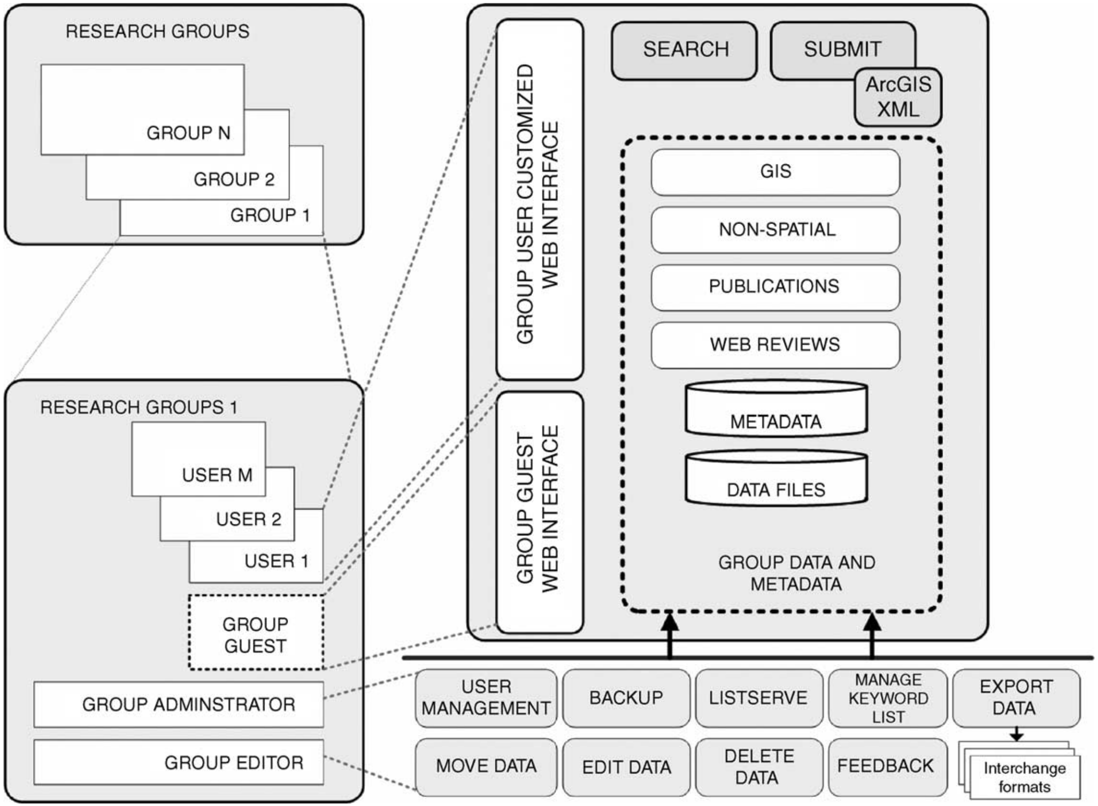


Fig. 3. The architecture of the Orsprivate.org system. Research groups are initially defined in the system by the ORS system developers (top left). A “group administrator” userid is created for each research group. Group administrators (bottom of figure) can create or delete group users, move submitted 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”.

Interpretation of Fig. 3 is as follows. Orsprivate.org provides services for multiple groups (1, . . ., *N*). Each group is comprised of 1 or more (1, . . ., *M*) users who gain access through a private user ID that is established by their group's local Orsprivate.org administrator. Group users access the system through a customized web interface with a homepage that is tailored to interface seamlessly with any other group web page that already exists for their project. Through their logging in, the system retrieves only Orsprivate.org metadata records that are associated with their research group. Other group records are unavailable.

The Intranet nature of the Orsprivate.org system requires additional functions for managing individual research groups. Just as the administrator functions in the public Open-Research.org system, group administrators act for the group as "editors" who not only review and manage submitted content (e.g., move submissions from the post to the search databases) as well as group access and communication. The bottom of Fig. 3 shows these additional functions that include a user management facility where group administrators can add, modify or delete group user accounts and monitor group activity statistics. In addition there are functions to communicate with a group through a server-supported listserv, functions to manage group keywords used as "pick lists" when submitting new metadata, and a mechanism to export group data to other formats (e.g., ASCII, MS Word, Excel, etc.) for a variety of possible reasons. These local administrator functions are web-based, allowing any group to have an administrator located anywhere geographically where there is Internet access.

Orsprivate.org provides an additional capability for very large research groups who are hierarchically organized into smaller units to address particular research questions. In the case of the BES research group, a group with over one hundred research participants scattered across the U.S. east coast, there are "subgroups" working on specific project elements. These tend to be broken down by academic disciplines, such as the "BES demography-socioeconomic" subgroup. In these instances, there are smaller teams of researchers who wish to share data only with others in their subgroup, with the idea that as data and other research products are ready to be shared with the broader interdisciplinary BES group they can be "promoted" to that higher, larger group (e.g., the all-BES group). The "move data" function at the bottom of Fig. 3 captures both the regular task of reviewing and moving submitted data from the post database to the search database, as well as the additional, but less frequent task of moving or copying a metadata record from a subgroup database to a related parent research group database.

The advance perhaps most appreciated by the Orsprivate.org users is the import GIS metadata function (Fig. 3), which allows a user to upload the XML metadata maintained within the ArcGIS^{TM} software package to the Orsprivate.org database. First, this function allows GIS analysts to submit the metadata in their organization's native GIS system, and yet still to make it available to the broader group or the public through the Orsprivate.org system. Second, it enables the GIS analyst to retrieve metadata in a native GIS format. This capability eliminates the burden of entering metadata from the ArcGIS system into the ORS system and, conversely, entering metadata from the ORS system back into the ArcGIS system. This function improves metadata quality, reduces the burden of submitting metadata to ORS and entering metadata into ArcGIS, and enhances user participation. Finally, in some instances, private research groups may be interested in sharing some of their data and other group content with a broader public. In the most recent release of this system, there is a “login as guest” option on the customized group guest (web) interface shown in Fig. 3. When entering metadata, group members can specify that the record being submitted be made “public”. Guest users can only access the database search components of Orsprivate.org and can only view records with this “public” designation.

3. Results

The public Open-Research.org system has been in operation since fall 2001 and the Intranet version, Orsprivate.org, has been in operation since early 2003. Five geographically distributed research groups that include participants from universities, government agencies, private institutions and NGOs have utilized the system, with three other groups contemplating using it. Primary participants come from the USDA Forest Service and the BES LTER. As of April 2004, the total number of metadata records in the Open-Research.org system was 2140. These break down into fourteen GIS entries, three non-spatial metadata entries, 2121 publication entries, and two web-reviews. Part of the reason for the large number of publication entries is an added function whereby Endnote^{TM} bibliographies can be imported into the system to share with others. But overall, this represents a relatively low rate of use.

As of April 2004, the metadata database in the Intranet version, Orsprivate.org currently contains a higher number—2306 records in total—which breaks down into 101 GIS, 27 non-spatial, and 2178 publication records, with most of the activity coming from the BES LTER and its associated research subgroups.

In retrospect, even the Intranet system has not been utilized as had been anticipated. Understanding why is important not only for the project itself, but so other projects and organizations that seek to develop Internet-based collaboration systems for file and research sharing can learn from the ORS experience. The question is: Why have some groups only partially embraced this system or stopped using it completely, given they have a strong need for a system to assist in their collaboration with geographically distant colleagues?

To shed light on this question, we conducted a survey of the five existing ORS “local administrators”. Because of the prominent role local administrators play in maintaining their group’s meta-database, it was important to get their assessment of ORS activity within their group. They act as the interface between the system and their set of end-users and are the ones that witness group activity on a day-to-day basis. Four open-ended questions were asked:

(1) When you first considered ORS for use with your group, what were your original goals?
(2) Have you found the system to be useful to you and your group? Why or why not?
(3) Did you encounter any barriers in system functions that limited the utility of the system? If so, what were they?
(4) What features might you like to see in a next version of this system that would really help your group?

In addition to this survey of administrators, we also held informal interviews with some of the ORS end-users. These dialogs revealed two common themes described below. We also introduce two additional themes—for a total of four themes—based on our own assessment and experience managing the two ORS systems.

3.1. A mismatch of system functionality with user needs

Several of the groups who continue to use ORS do so because they use it for its intended purpose—the sharing of metadata about relatively complete and stable datasets and research products (e.g., white papers, etc.). Several of the groups who started using ORS, and then reduced their use over time, tried to use it initially for a purpose that was not in our original design: storage of a rapidly changing document that all users were working on. The groups that reduced their activity found that email with file attachment was an easier method for collaboration.

3.2. Importance of system ease of use and speed

One group reported running into initial accessibility problems for team members during the early release days of the Orsprivate.org system. Some of these problems were the result of identified bugs in the system, but other troubles could be attributed to the need for additional local administrator or team member training beyond what was done over the phone. In addition, several teams wanted to share extremely large files (e.g., multispectral satellite images or digital photographs) and depending on their Internet connection, transfer speeds could be slow. In these instances, users were encouraged to complete metadata on the system so that others were aware that their datasets exist, but *not* to upload their large files. Users were then encouraged to keep their data on another secondary storage device (e.g., a CD) and have people contact them if they want a copy. Until Internet transmission speeds become faster and more universal, this procedure of filing metadata coupled with local CDROM or other backup storage will continue to be important.

3.3. Preference for Intranet systems

Once Orsprivate.org went on line, it became clear that users preferred the Intranet system over the publicly accessible Open-Research.org system. As of April 2004, activity on Open-Research.org has nearly ceased, whereas activity on Orsprivate.org continues to gain momentum, particularly in the case of the BES group. And several other new groups have requested use of the system. Still, while the emphasis is on private sharing of data, there continues to be interest in providing the option to share with the broader public when the group is ready to do so. This is why Orsprivate.org now has the “login as guest” feature to allow non-group visitors to search “public” designated records in the group database.

3.4. The challenge of metadata: workplace culture and habits

The task of entering metadata is something that people appear to resist unless it is mandated in their job descriptions or through other enforcement mechanisms. The goal was to develop metadata forms that followed the standards developed by the Federal Geographic Data Commission for GIS data, and to develop similar metadata structures for non-geographic and publication data. By using the USGS’ PC-based “MetaLite” system with its minimum metadata standards as a guiding template, the attempt was to design a set of forms that made it as easy as possible for the end-user and still conform to these standards. Several meetings were held early on with users to review these fields from a usability standpoint and to resolve any confusion over the meaning of some fields.

Even with all these measures taken to make it easy for the end-user, it became apparent that there was still significant resistance by group members to take the time to enter their metadata. There are several explanations for this. Most people are extremely busy in their jobs and although dataset documentation is acknowledged as something that is important, for many it is at the low end of the daily to-do list in practice. When a dataset owner does take the time to undertake this work, the task often involves inventorying legacy datasets that predate metadata tools and/or lack written metadata, the metadata therefore are not easily reconstructed.

4. Conclusions: implications and future directions

From these experiences, there are two areas of insight that should be of general interest to readers. First are the implications of this project to practical issues of managing data in an Internet world. Second are the implications of this project for the emerging trend of open content-based collaboration.

4.1. Implications for managing data in an Internet world

There is a growing trend toward the development of Internet-based collaboration tools, and the experience of the ORS project may provide useful insight for the design of next generation systems. The variety of collaborative systems listed in Table 1 supports the contention that "virtual group" collaboration systems are an important emerging area of software development. The two ORS systems focus more heavily on data management issues, whereas these other systems in Table 1, with the exception of SourceForge, focus more attention on synchronous and asynchronous communication and collaboration tools and in project management applications. We expect that many of these systems will be moving toward providing many or all of the functions listed in Table 1, and therefore the lessons here related to data management will be of interest to others working on the next generation of these types of systems.

4.1.1. The "science" of filing

The first lesson gained from this project is that while the ORS is a "virtual" file cabinet, it is still a filing cabinet. What are the implications of this observation? We see at least three. First, there is an underlying "science" to organizing documents and developing good habits for keeping a filing system functioning well. Many team members may not possess these important skills.

Second, long-term projects will have a higher need than short-term projects for establishing good systems of filing and encouraging good filing habits by their team members. Therefore, long-term projects with geographically distributed teams will be more interested in using systems like ORS and short-term projects will more likely utilize tools like email and attachments as their system for file sharing. This reflects our empirical observations.

Third, it appears that it takes time for members of long-term projects to stop behaving as if they are working on short-term projects. Our hypothesis when the project started was that the adoption rate of a system like ORS and its attendant new filing approaches and habits would resemble an exponential curve: slow at first and then faster over time. Our expectations are in line with and supported by theoretical work on innovation diffusion, which suggests that a critical mass must be established in order for an innovation to become self-sustaining (Rogers, 1995). In some instances, it could take years before a critical mass can be achieved and for organizational culture to adopt new filing practices without added and strong incentives to force change (discussed below).

4.1.2. Incentives and penalties for encouraging good filing procedures

Changing filing habits in organizations or virtual organizations may require incentives and penalties for non-compliance. In some instances, the incentives and penalties for using a system like ORS are well established and quite clear. For example, in the context of the LTER program, if a research group fails to complete metadata for their project, then the information management component of their renewal review will be negative, and this could jeopardize their future funding.

In other instances, incentives are established already for some aspects of data management, but specific penalties may be lacking. For example, in the USDA Forest Service, scientists undergo periodic performance reviews in a process called “panels” (USDA Forest Service, 1995). One area of review is whether the scientist has produced documented, digital databases over the course of the year. This procedure rewards the production and documentation of such databases (e.g., through possible promotion, etc.), but the procedure does not specify any penalty if a scientist has failed to document his or her database.

The above are examples of existing incentives and penalties, but there are other new possibilities that, to our knowledge, have yet to be tried. Through systems like ORS, it is possible to use program tracking mechanisms that help to gauge the utility and/or significance of content posted and made public. In other words, just as the *Web of Science* (http://www.isinet.com/isi/) allows a researcher to track how many times a publication has been cited, it is possible to provide mechanisms (such as follow-up surveys to people who downloaded particular datasets) that would track how many times data have been accessed or downloaded and provide ways to determine whether the data has contributed to some significant research findings. Organizations such as the USDA Forest Service already have mechanisms to do such tracking (e.g., through tallies of data requests or mailings of CDs) and give the creators of such data credit for the number of times some data have been used. For example, in their performance reviews, USDA Forest Service scientists are required to provide documentation on the dissemination of electronic or audio products they have produced (USDA Forest Service, 1995). They receive higher performance marks when they can show that their products have been widely disseminated. Programming such mechanisms into online systems like ORS extends some of these existing manual practices. Also, funding agencies like the NSF consider it important for major research groups, like LTER, to know how many times data have been downloaded from LTER websites.

4.2. Future ORS directions

In addition to continued system maintenance and support of current and interested new groups, there are several possible future directions for the ORS systems (Fig. 4). These enhancements fall under four categories: (1) decommissioning parts of the system that are underutilized; (2) improving the ability for interested parties to find, over the Internet, ORS data that are published for public consumption; (3) developing mechanisms to measure the utility of this public data; and (4) enhancing the system so that it more effectively supports “open source” and “open content” collaboration. This last option could lead to a new paradigm for the way scientific research in agriculture and other fields are conducted.

4.2.1. Decommissioning the original Open-Research.org system?

Given the degree to which Orsprivate.org is now used compared to the original Open-Research.org system, it is possible that over this year we may decide to decommission the Open-research system and maintain and enhance only the Orsprivate.org system. In short, from a user group perspective, an Intranet system with “guest” access appears to be a design preferable to an entirely open access system.

4.2.2. Improving the ability for parties to find ORS data: linking to the National Spatial Data Infrastructure network and to web search engines

In general, most research groups are interested in making data and other research products widely available after a period for “private analysis” and publication of these data. Some organizations are mandated by their funding agencies (e.g., the NSF) or broader organizational policies (e.g., the U.S. Department of Agriculture Forest Service) to make their data public after some reasonable amount of time.

One enhancement to Orsprivate.org currently being considered is to make it a server associated with the National Geospatial Data Clearinghouse network (http://clearinghouse1.fgdc.gov/) so that group geographic data designated “for public consumption” can be searched and found through the National Spatial Data Infrastructure (NSDI) web search facility. To do this, a program would be written that would automatically run at regular intervals to copy or link publicly available records in Orsprivate.org and reformat them to the standards required for NSDI servers (see http://www.fgdc.gov/publications/documents/clearinghouse/chouse.pdf for more information about how to participate in this network).

But this would make only the public GIS data stored in Orsprivate.org more accessible to Internet users. How might people unaware of the Orsprivate.org system find the other sorts of metadata and data (e.g., publications, non-spatial data) contained within the system? The problem of how to facilitate guest access to the Orsprivate.org database is related to what has been called the “deep” or “invisible web” (Sherman and Price, 2001). It is fast becoming the norm for people to turn to web-based search engines like Google$^{\mathrm{TM}}$ as their first step toward finding information (Boyle, 2004), but the “deep web” of private databases is often invisible to Google$^{\mathrm{TM}}$ searches. How can that “invisible web” content be made more accessible to the web search engines like Google$^{\mathrm{TM}}$ that have become so popular? To make this connection, Orsprivate.org has a function that allows local administrators to extract or export metadata from the Orsprivate.org database for direct posting on group web

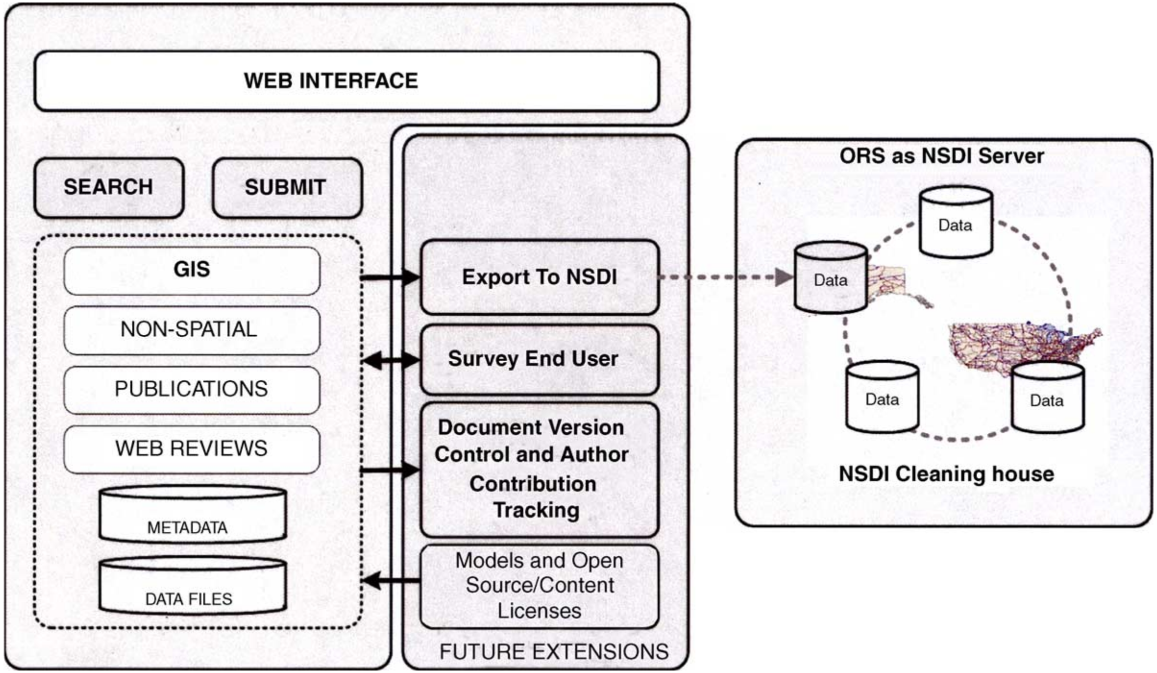


Fig. 4. Possible future extensions to Orsprivate.org functionality that are currently being discussed. The “survey end user” has become the immediate priority, for several groups are interested in understanding the impact that their data and other documentation have on a broader community outside of their group. In addition, the existing search and submit functions are being enhanced to provide “plug-in” web services that can be added to an organization’s web site, rather than having to go to the Orsprivate.org site.

pages that web crawler programs will find (Fig. 3). Over the next year some systems of evaluation will be established to better understand if this strategy (discussed more in the following section) is effective.

4.2.3. Developing a "survey of end users" available at the downloading data function

One of the major research questions facing the BES and is common to many research groups is "to what extent are the data and research being produced making a difference and being utilized by various researchers external to the group, as well as to other stakeholders such as public officials, non-profit groups and other interested citizens?" The impact of their research products is something most if not all research groups are interested in understanding more fully.

The traditional way of documenting that kind of utilization is by recording the number of requests of external parties for research products (i.e., reprints of papers). This, for example, is one way that the USDA Forest Service evaluates the impact of their research. The interactivity the web brings provides new opportunities for studying the interest in information produced by a research group. Over the next year tracking information will be developed into the Orsprivate.org database by adding a counter that records the number of times any particular dataset or research product is downloaded. An (optional) online survey instrument will be developed asking about who the person is, what their particular interest is in the data or information being downloaded, and whether they give ORS personnel permission to contact them in the future to see how useful the data or information actually were for their work.

4.2.4. ORS "plug-ins" and web services

There have been concerns voiced by some users of ORS that by having it be a separate website it takes away attention from their own website. In other words, ORS users worry about linking to another website that may take away from their own world-wide web identity. For this reason, for Orsprivate.org, version 3.0, we are moving toward a "plug-in" kind of design, where, a research group can include the "submit" and "search" functions directly into pages of their own web system, rather than invoking standard pages residing on the Orsprivate site. We intend to do this using the Extensible Markup Language (XML), Extensible Stylesheet Language Transformations (XSLT), ColdFusion^{TM} forms, and the Verity^{TM} engine.

4.2.5. Moving toward systems that support "open source/content" scientific collaboration

In Section 1, we stated that the model set forth by open source programming projects was an early driver of the concept of the ORS system (Schweik and Grove, 2000; Schweik and Semenov, 2003). Over the last 4 years these ideas have evolved and we now see this as an area of real potential for future cross-organizational research collaboration. Systems like ORS (with further enhancements) will provide the infrastructure to support such collaborative work.

Put simply, open source programming projects are collaborative, Internet-based endeavors, where the source code of programs are freely readable by others. The primary innovation that drives open source projects is the licensing. While there are many licensing variants, the first, and perhaps most dominant license used in the open source software domain is the "GNU General Public License" or "GPL" developed by Richard Stallman and colleagues back in 1984 (Perens, 1999). Licenses such as the GPL are innovative in that they change the rules of the game in the way software is distributed and further developed. Rather than fully copyrighting (e.g., "all rights reserved" with distribution of unreadable compiled proprietary software), these licenses often include rules requiring that the readable source code (and compiled versions) can be freely distributed, and in many cases that users or other developers can freely contribute new enhancements to this software. Moreover, these licenses often carry a "viral" component to them that stipulates that any new derivative work based on a particular software automatically carry the same license as its earlier variant.

Over the last decade, Internet-based infrastructure has been developed to enhance the distribution of open source software and to help coordinate sometimes nearly global teams of volunteer and paid programmers to contribute to the further development of open source code. One example of this type of collaboration system is the website Sourceforge.net, which advertises (as of July 2004) over 80,000 open source programming projects being "housed" in its repository. Some of the collaborative functionality of Sourceforge.net is shown in Table 1. Central components to support such collaboration include software version control and team communication facilities.

While there are several high profile open source success stories such as the Linux operating system and Apache Web Server, there are a large number of open source software initiatives that would probably be considered failures. And it has only been recently that serious scientific efforts have been undertaken to understand these types of collaborative systems (see, for example, the report by Ghosh et al., 2002 at http://www.infonomics.nl/FLOSS/report/index.htm). But movement of free sharing of intellectual property by individuals through alternative licensing schemes could be a milestone in the way humans collaborate in the future and has the potential to extend well beyond computer science (Bollier, 1999; Schweik and Grove, 2000; Stallman, 2001; Stadler and Hirsch, 2002; Weber, 2004).

It should also be noted that the idea of open source over the Internet has the potential to greatly speed up our ability to advance scientific inquiry by making information really "open" and readily available on the Internet, rather than being slowed down because of licensing barriers that restrict access to information except for people who have the ability to pay for it (Lucky, 2000). For an "existence proof" to support this statement, one only has to reflect on the growth of the web from 1994 to 2000. While not formally open source licensed, web pages were, by default, open source. The reason the web grew so exponentially during this time was because of the "view source" options available in web browsers such as Internet Explorer and Netscape. These options allowed those who had web access to read others' HTML code, learn from it, and incorporate what they learned in their own web pages. It was this open source feature of the web that drove the rapid expansion of the web globally (Lessig, 2001).

Since our initial realization in 2000 of the importance of these open source principles for broader scientific communication, new forms of open source licenses have emerged for other forms of intellectual property that are not software. Intellectual property scholar Lawrence Lessig and others working at CreativeCommons.org have developed 11 "open content" licenses that allow authors of any type of intellectual property (e.g., music, text, etc.) to copyright some, but not all rights. In other words, developers of new ideas are not limited to the binary choice of either choosing full copyright (e.g., “all rights reserved”) or making the content “public domain”. The CreativeCommons.org licenses allow authors to choose which rights they wish to preserve and which rights they wish to relinquish (Stix, 2003; see also http://creativecommons.org/learn/licenses/).

While it is too early to determine, this model of “Open Content” licensing of intellectual property, coupled with Internet-based systems of metadata and data or content sharing like ORS, could lead to important new forms of scientific collaboration that could extend beyond organizational lines and span the globe. And there are other open content initiatives currently underway (see Schweik and Semenov, 2003 for a list of some such projects). Our own efforts currently are trying to apply these ideas to promote an open source/content research collaboration in the context of landuse change modeling (Schweik and Semenov, 2003). This means ORS (or a comparable system) would require the capacity to allow the submission of landuse models and submodels (e.g., in potentially a variety of forms and approaches such as statistical models, geographic information system-based models, etc.) as well as other content such as model documentation and distance learning material, papers describing applications of these models, etc. All of these components would need to have some form of open content license attached to them, in addition to systems for tracking intellectual property contributions as well as version control (Fig. 4).

A central consideration for getting scientists to contribute their intellectual property to such a system is incentives. In many research areas, the central incentive for scientists is to be able to publish their intellectual property so that they can cite this publication in their CV. Formal peer-reviewed publishing is the central incentive for academics and many other scientists employed by research institutions. This means that next generation systems like ORS or Sourceforge.net, supporting scientific research (not computer programming, although there will be a convergence of these, we believe) will have to move toward new forms of “e-journals” that embrace both open content licenses and include peer-review and formal publishing that move beyond what is done in standard journals today and most e-journals. In other words, through the convergence of open content licensing, systems like ORS, and the idea of peer-reviewed e-journals, there is the opportunity to move beyond the publishing of final results and to the publishing of all sorts of research products that lead to final results (e.g., data, intermediary research steps, etc.). It is this convergence that could lead to a whole new paradigm in the way science is conducted in the future. Systems like ORS, with improved development and enhancement, could contribute significantly toward new forms of scientific collaboration.

Acknowledgements

We gratefully acknowledge support provided by the USDA Forest Service’s Burlington Laboratory (4454) and the Northern Global Change Research Program under Cooperative Agreement No. 11242343-063-A1. Some components of this work have also been sponsored by the Cooperative State Research Extension, Education Service, U.S. Department of Agriculture, Massachusetts Agricultural Experiment Station, under Project No.

MAS00847, as well as by the Center for Public Policy and Administration at the University of Massachusetts, Amherst, and the Department of Natural Resources Conservation at the University of Massachusetts, Amherst. We also thank Robert and Joanne Schweik, two anonymous reviewers and the journal editors for very helpful comments on an earlier version of the manuscript.

References

Ancona, D., Caldwell, D., 1990. Information technology and work groups: the case of new product teams. In: Galegher, J., Kraut, R., Egido, E. (Eds.), Intellectual Teamwork. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 173–190.

Bollier, D., 1999. The power of openness: why citizens, education, government and business should care about the coming revolution in open source code software. http://h2oproject.law.harvard.edu/opencode/h2o/.

Boyle, J., 2004. Mertonianism unbound? Imagining free, decentralized access to most cultural and scientific material. In: Paper Presented at the Workshop on Scholarly Communication and an Information Commons, Indiana University, Bloomington, April 1.

Burnham, J.C., 1990. The evolution of editorial peer review. J. Am. Med. Assoc. 263 (10), 1323–1329.

Cohen, J., 1996. Computer mediated communication and publication productivity among faculty. Internet Res. 6, 41–43.

Estrin, G., 2000. Computer network-based scientific collaboration in the energy research community, 1973–1977: a memoir. IEEE Ann. Hist. Comput. 22, 2.

Finholt, T., Sproull, L., Kiesler, S., 1990. Communication and performance in ad hoc task groups. In: Galegher, J., Kraut, R., Egido, E. (Eds.), Intellectual Teamwork. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 291–325.

Ghosh, R.A., Robles, G., Glott, R., 2002. Free/Libre and Open Source Software: Survey and Study. Technical Report. International Institute of Infonomics. University of Maastricht, The Netherlands, http://www.infonomics.nl/FLOSS/report/index.htm.

Gillman, D.W., Appel, M.V., LaPlant Jr., W.P., 1996. Design principles for a unified statistical data/metadata system. In: Proceedings of the 8th International Conference on Scientific and Statistical Database Management, Institute of Electrical and Electronics Engineers.

Hesse, B.W., Sproull, L.S., Kiesler, S.B., Walsh, J., 1993. Returns to science: computer networks in oceanography. Commun. ACM 36 (8), 90–101.

Johns, A., 2001. The birth of scientific reading. Nature 409, 287–289.

Kerschberg, L., Gomaa, H., Menasc, D., Yoon, J.P., 1996. Data and information architectures for large-scale distributed data intensive information systems. In: Proceedings of the 8th International Conference on Scientific and Statistical Database Management, Institute of Electrical and Electronics Engineers.

Kronick, D., 1990. Peer review in 18th century scientific journalism. J. Am. Med. Assoc. 263 (10), 1321–1322.

Lessig, L., 2001. The Future of Ideas. Random House, New York.

Lipnack, J., Stamps, J., 1997. Virtual Teams. Wiley, New York.

Lucky, R., 2000. The quickening of science communication. Science 289 (5477), 259–264.

Moen, W.E., 2001. The metadata approach to accessing government information. Government Inform. Quart. 18, 155–165.

Perens, B., 1999. The open source definition. In: DiBona, C., Ockman, S., Stone, M. (Eds.), Open Sources: Voices from the Open Source Revolution. O'Reilly and Associates, Sebastopol, CA.

Rogers, E., 1995. Diffusion of Innovations. Free Press, New York.

Schweik, C.M., Grove, J.M., 2000. Fostering open-source research via a world wide web system. Publ. Admin. Manage.: Interact. J. 5, 3, http://www.pamij.com/5_4/5_4_2_opensource.html.

Schweik, C.M., Semenov, A., 2003. The institutional design of open source programming: implications for addressing complex public policy and management problems. First Monday 8, 1. http://www.firstmonday.org/issues/issue8_1/schweik/.

Sen, A., 2004. Metadata management: past, present and future. Decis. Support Syst. 37, 151–173.

Sepic, R., Kase, K., 2002. The national biological information infrastructure as an e-government tool. Government Inform. Quart. 19, 407–424.

Sherman, C., Price, G., 2001. The Invisible Web: Uncovering Information Sources Search Engines Can't See. CyberAge Books, Medford, NJ.

Sproull, L., Kiesler, S., 1991. Connections. MIT Press, Cambridge, MA.

Stadler, F., Hirsch, J., 2002. Open source intelligence. First Monday 7, 6. http://www.firstmonday.org/issues/issue7_6/stalder/.

Stallman, R.M., 2001. The free universal encyclopedia and learning resource. In: Werry, C., Mobray, M. (Eds.), Online Communities: Commerce, Community Action, and the Virtual University. Prentice-Hall, Upper Saddle River, NJ, pp. 257–269.

Stix, G., 2003. Some rights reserved. Sci. Am., 10.

USDA Forest Service, 1995. A Guide for Preparing Research Scientist Position Descriptions. USDA Forest Service, Washington, DC.

Walsh, J., Bayma, T., 1996. Computer networks and scientific work. Soc. Stud. Sci. 26, 661–703.

Weber, S., 2004. The Success of Open Source. Harvard University Press, Cambridge, MA.

Ziman, J., 1969. Information, communication, knowledge. Nature 224, 318–324.