Digitizing Rangelands: Providing Open Access to the Archives of Society for Range Management Journals

Title: Digitizing Rangelands: Providing Open Access to the Archives of Society for Range Management Journals

By Yan Han, Jeanne Pfander, and Marianne Stowell Bracke

Keywords:
Rangelands, open access, digitization, metadata, MARC, Dublin Core, electronic journals, digital library system, web log analysis

Abstract:

The University of Arizona Library is a vital participant in the AgNIC Rangelands project and has contributed to the initiative in many ways. For example, in the mid-to-late 1990's the Library digitized the backfiles of the *Journal of Range Management* for open access on the web. Funding and completing digitization projects such as this is a complicated process and requires many decisions along the way. This paper outlines the process taken by the Library to manage a more recent project to scan back issues of the journal *Rangelands*. It starts with the decision to develop a project plan and request funding from the National Agricultural Library (NAL). It continues on to describe negotiating agreements with project partners, the process for outsourcing of scanning, the design of the technical infrastructure needed to support digitized content, and issues of sustainability that any digital library encounters.

Introduction

The University of Arizona (UA) has been a part of AgNIC (http://www.agnic.org), centered at the National Agricultural Library, since its inception in 1995 (Gardner, Gilbertson, and Hutchinson 2002). UA's area of responsibility is rangelands. Rangelands are crucial to Arizona, but are also of global importance as almost 50% of the world's landmass is rangelands. The Arizona AgNIC rangelands project team is interdisciplinary, composed of librarians, technical staff, and rangeland experts from several collaborating units in the University's College of Agriculture and Life Sciences and the Library. (Hutchinson and Ruyle 2002) The mission of the Arizona AgNIC project is "to provide electronic access to the full scope of information in the field of rangelands and rangeland management to people everywhere and of all knowledge levels by collecting, creating, evaluating, and organizing relevant resources." The Arizona Rangelands site resides at http://ag.arizona.edu/agnic/az/index.html.

AgNIC Rangelands has continued to evolve. In 2002, the Arizona AgNIC team brought together Cooperative Extension agents and librarians from Land Grant institutions from 12 western U.S. states to expand the AgNIC Rangelands project (Hutchinson, Jones, and Ruyle 2003). This initiative, which has since grown to 21 participating institutions in 19 states, is now known as the Western Rangelands Partnership, and a central web site has been created named "Rangelands West" (http://rangelandswest.org). From this site, links lead to state pages that address local needs and issues.

History of Arizona AgNIC Journal Scanning

The Arizona AgNIC Rangelands team has a long history of pursuing opportunities to create or support the creation of online access to more content. In 1995, two events were taking place that came together serendipitously. The Society for Range Management (SRM), the professional scientific society and conservation organization that studies the varied resources of the rangelands, contacted NAL about the possibility of digitizing backfiles of the *Journal of Range Management* (*JRM*). The *JRM* (recently renamed *Rangeland Ecology and Management*) is a peer-reviewed journal published by SRM. It is a key journal for the study of rangelands and covers all aspects of rangelands management and research. Simultaneously, the Arizona AgNIC project was just getting off the ground and was beginning to add content to the site. NAL brought the two groups together. An agreement was reached for the University of Arizona to digitize the backfiles of the JRM and Memorandum of Understanding (MOU) between the UA, SRM and NAL was drafted. NAL provided some funds for the digitization work. SRM agreed to provide the print copies of the back issues from 1948 – 1994 for the Arizona rangelands project team to scan. The process of scanning, plus creating the web architecture for delivering the initial online backfiles, was done in-house at the UA Library and was completed in 1999 (Hutchinson and Ruyle 2000). Two years later, the project team added volumes for 1995 – 1998 which were provided in digital form by Cornell University as a by-product of TEEAL (The Essential Electronic Agricultural

Library - http://teeal.cornell.edu/). More recently, additional volumes for 1999-2001 have been provided by SRM in digital form. SRM will deliver additional volumes as the rolling two year "window" agreed upon in the MOU continues to move. The JRM archive site (http://jrm.library.arizona.edu/) includes the full text for all issues of the journal from 1948 – 2001. Issues can be accessed by browsing through an index by date. For the years 1983 – 2001, the archives can also be searched by general keyword, author, or title.

Digitizing *Rangelands* - Funding Opportunity & Writing the Proposal

In April 2004, NAL announced that funding for small cooperative agreements was being made available to AgNIC partners. The objective of the cooperative agreement was "to support an effort between participating institutions and NAL to build full-text content which would deliver information to Internet users through the Agriculture Network Information Center (AgNIC)." (Gardner 2004) The deadline for proposals was June 15, 2004, with funding to become available on October 1, 2004 for the fiscal year ending September 30, 2005.

For several years, the Arizona Rangelands AgNIC group had been discussing the desirability of working with SRM to make the full-text content of the backfiles of the journal *Rangelands* available online. This had never moved past the idea stage due to lack of funding from either SRM or UA. With the opportunity presented by the NAL Cooperative Agreement funds, the authors of this paper met in May 2004 to draft a proposal.

Initially, we considered seeking funds for either the digitization of *Rangelands* backfiles (vols.1-20, 1979-1998) or a project to develop indexing and metadata for the *Journal of Range Management* articles between the years 1948-1982 (vols.1-35) to add to the JRM Archives database (and possibly to the AGRICOLA database). NAL/Agricola has only fully-indexed JRM since 1983, although there was some cursory "free choice 'title enhancement'" indexing for JRM from 1970-1983. It quickly became apparent that the indexing project would be too expensive given the amount of funding that NAL had available to distribute. We decided therefore to focus on the digitization of the *Rangelands* backfiles.

The proposal outline included the following: objective, approach, statement of mutual interest and budget. The wording for the objective was supplied as a template by NAL with only a few key phrases to be inserted by the requesting institution. The "approach" section was the work plan for the project – "How will it be done? What are the milestones? How often will the institution provide a status report to NAL?" The statement of mutual interest was a description of how the project is beneficial to each institution. The budget, of course, could include salaries and wages, fringe benefits, materials and supplies, travel, publication costs and indirect costs.

The following "approach" and work plan was developed:

The University of Arizona Library will digitize 20 volumes (v.1-20, 1979-1998) of the journal *Rangelands*…

The UA Library will arrange for the outsourcing of the digitization of the issues to a vendor and the subsequent mounting of the files on the Library's server. It should take 6-8 months from the distribution of the funds to complete this project. The UA Library will provide a status report and/or final report to NAL at the 6 month mark. The work plan includes the following steps:

1. UA Library works out memorandum of understanding with SRM
2. UA Library develops contract with the vendor, including standards and quality requirements
3. SRM ships *Rangelands* volumes to the vendor
4. Vendor digitizes volumes
5. Vendor delivers digitized files (PDF) on CD-ROM to UA Library
6. UA Library ensures that vendor has met quality standards
7. UA Library creates template for metadata entry
8. UA Library inputs metadata and/or downloads metadata from Agricola
9. UA Library loads files into the system
10. UA Library creates a front page and graphical design for the data
11. UA Library creates additional guides (e.g., FAQ)
12. UA Library tests system
13. UA Library provides copies of files (PDF) on CD to SRM and NAL
14. UA Library and SRM market newly available online *Rangelands* content
15. UA Library will provide a project report at the end of the project and will report to NAL bi-monthly by telephone on the status of the project.

In addition, we "cost-ed out" the amount of time each of the three project team members would be contributing at each stage of the project. This fed into the development of the budget. The UA project was approved for funding by NAL, along with other AgNIC projects.

Writing the Memorandum of Understanding (MOU)

Following the precedent established by the *JRM* digitization project, we created a MOU between the University of Arizona and the Society for Range Management regarding the digitization of the journal *Rangelands*. Responsibilities for SRM and UA were defined.

SRM agreed to provide a) print copies of volumes and b) rights to unlimited, free-of-charge use of the digitized data consistent with Title 17 of the U.S. Code on Copyright.

UA Library agreed to a) scan the printed material and create digitized TIFF and PDF files; b) make this information available over their World Wide Web (WWW) site; c) provide SRM with 2 free copies of the TIFF and PDF files and make additional copies at cost if requested; d) make the Rangelands information from the UAL web site directly

accessible from other web sites that emphasize rangeland information, particularly the SRM home page site (http://www.rangelands.org/).

Negotiations for the MOU ended up taking significantly more time than we hoped or expected due to some staff changes in the SRM offices and communication delays between the UA Library and the UA Contracts office.

RFP and Vendor Agreement

A Request for Proposal (RFP) was drafted to describe the context and requirements, including scanning quality and quantity, file formats, a file naming conversion and delivery methods. A consistent and meaningful file naming convention is critical for both systems and users, because it provides system compatibility, interoperability across platforms and file relationship information. The naming convention was finalized based on standards (ISO-9660 for a standard CD-ROM file system) and best practices. For example, a file name "rangelands_v2_n2_51-52.pdf" indicates that this is a PDF article from journal *Rangelands* volume 2 no. 2 from page 51 to page 52. The vendor was selected and a contract was signed between the vendor and the UA Library for the requirements, costs, and timing for completion of the work. When the contract was in place, SRM shipped the *Rangelands* backfile volumes to the vendor for digitization.

Developing the Infrastructure

Internet search engines such as Yahoo and Google have created high user expectations regarding access to information. Digital information must be connected and accessible anytime and anywhere. A strong infrastructure (technical and non-technical) that delivers accessible digital information is critical in the networked information environment.

In agreeing to digitize the backfiles and provide open access for the journal *Rangelands*, it was important that we first define the non-technical infrastructure for this "digital library". The MOU establishes the intellectual property rights and allows the UA Library to deliver the journal content online.

The technical infrastructure consists of digital objects, associated metadata, and digital library system. Good digital objects principles involve production format, persistent identifier, and associated metadata (NISO Framework Advisory Group 2004).

Digital Objects

A good production format must be chosen to support the divergence of customers' computing environments for both current and future access. PDF, JPEG and HTML are the most predominant access formats for journal materials, while TIFF is the main preservation format. The current best practice recommends that for textual materials such as journal articles, objects in 150 – 300 dpi with 1 bit black-and-white or 8 bit grayscale or 24 bit color in PDF / JPEG are good for access; while objects in 300 - 600

DPI with 1-bit black-and-white or 8 bit grayscale or 24 bit color in TIFF are suitable for preservation.(Western States Digital Standards Group 2003) (Seadle 1998)(Berger 1999) (Fleischhauer 1998)

The UA Library project team scanned a few sample objects in multiple configurations to evaluate document quality and download speed regarding file size. The tests showed that TIFF objects in 300 dpi with 8 bit grayscale appropriately balance cost, access, and preservation, while 35% compression of the correspondent PDFs gives smaller file sizes with good quality. For example, a 35% compression in PDF of an 8-page article has the file size of 4M bytes, which is appropriate for customers with different connection speeds. When scanning an article, typical scanning software will generate an object in TIFF format and with an Optical Character Recognition (OCR) module the software can also produce an OCR'd PDF file.

Based on collection requirements and research, objects saved in PDF format (with 35% compression in 300dpi with 8 bit grayscale) have been chosen as for access, while objects in TIFF and PDF format (in 300dpi with 8bit grayscale) will be used for digital preservation.

Metadata

Good metadata should be appropriate to the collection, use controlled vocabularies, and support interoperability (NISO Framework Advisory Group 2004). Index records from NAL for *Rangelands* are obviously the best choice with respect to quality and cost. Although these index records are in traditional MARC format, they are created by professional catalogers using a controlled vocabulary: the NAL Thesaurus. Tasks associated with metadata for the project included getting *Rangelands* index records from NAL and mapping them to a format suitable for our digital library system.

With help from NAL, 1220 records were sent, each of them corresponding to one *Rangelands* article. Since digital library systems usually require different metadata formats such as Dublin Core (DC) instead of MARC, a crosswalk (mapping) was necessary for transferring records in MARC to their counterparts in DC. Default crosswalks from MARC to Dublin Core were published by the Library of Congress in 2001(Library of Congress 2001), including MARC to DC (Unqualified) and MARC to DC (Qualified). It is worth noting that "not all possible MARC fields are included in this mapping, but only those considered useful for broad cross-domain resource discovery"(Library of Congress 2001). MarcEdit is the tool that most librarians use for editing MARC records, and it contains a XSLT script that provides MARC to DC (Unqualified) mapping. The original script provides simple mapping, but misses some important MARC fields such as 246 (alternative title), 300(physical description, 651(subject), and 773(host item entry). Field 651 provides important geographic information, while field 773 and field 300 contain critical relationship information about articles. These fields are vitally important for the *Rangelands* collection and will improve customers' experience in browse, search and discovery.

Table 1: a sample *Rangelands* record in MARC

100	1\	$aCoupland, J.W.
245	10	$aFrom truck to well puller.
246	0\	$iEnriched title:$aFrom truck to well puller [Source of water for livestock, New Mexico]
300	\\	$ap. 51-52.$bill.
651	\3	$aNew Mexico
700	1\	$aYarbrough, C.C.
700	1\	$aGarcia, E.L.
773	0\	$tRangelands.$gApr 1980. v. 2 (2)$x0 0190-0528$7nnas$9jnl30643
852	\\	$aDNAL$iSF85.A1R32

Table 1 is a sample MARC record. The 773 and 300 fields indicate the article is from April 1980 volume 2, no. 2 pp. 51-52, which is very important linkage information. A modification of the original script is required to map these critical fields. The authors examined and modified the XSLT script by mapping the following fields: 246 to <dc: title>, 651 to <dc: subject>, 773 $x to <dc: relation>, 773 $t and $g to <dc: relation>, and 300 $a to <dc: description>. The new script generates more information for a record, maintaining these critical relationship and subject information for the system and customers. Table 2 is the result of mapping the above sample MARC record (Table 1) to DC using the new script.

Table 2: a sample *Rangelands* record in DC

```
<rdf:Description>
        <dc:title>From truck to well puller.</dc:title>
        <dc:title> Enriched title: From truck to well puller [Source of water for livestock, New
        Mexico] </dc:title>
        <dc:creator> Coupland, J.W. </dc:creator>
        <dc:creator> Yarbrough, C.C. </dc:creator>
        <dc:creator> Garcia, E.L. </dc:creator>
        <dc:type>text</dc:type>
        <dc:language>eng</dc:language>
        <dc:subject>New Mexico</dc:subject>
        <dc:relation>0 0190-0528</dc:relation>
        <dc:relation>Rangelands.</dc:relation>
        <dc:relation>Apr 1980. v. 2 (2)</dc:relation>
        <dc:description>p. 51-52. </dc:description>
</rdf:Description>
```

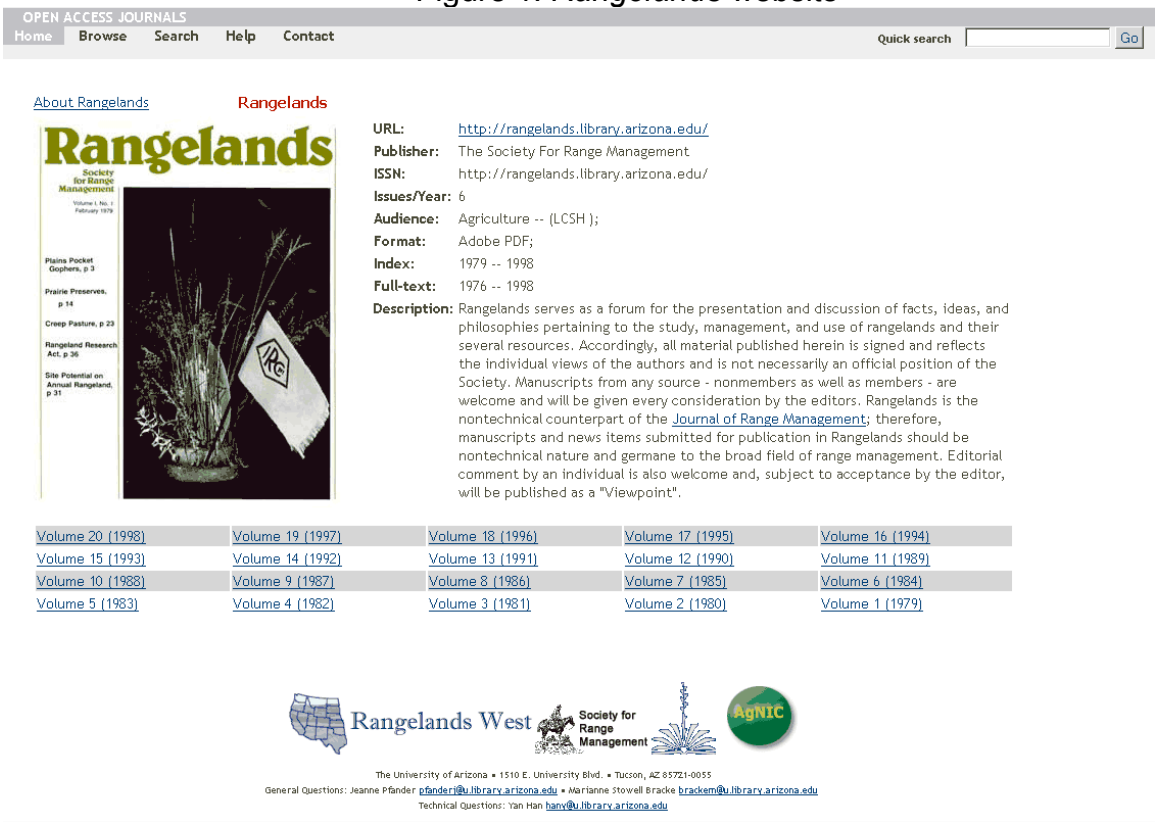
The Digital Library System

Usually a single server is used as the hardware platform for most digital library software. However, this approach cannot guarantee the availability of hardware, since there are some factors that can cause server failures, including hardware (e.g., hard drive breakdown), natural disasters (e.g., earthquake), network (e.g., router) and human error. To provide redundancy and ensure a high availability computing environment, a high availability computing platform based on a Storage Area Networks (SAN) was employed as the hardware platform. This computing platform provides a better solution than the traditional approach (Han 2005).

With respect to digital library software, the *JRM* Archives were set up in SiteSearch, currently maintained as an open source software. However, the divergence of the index records and their full-text documents across two different servers causes unnecessary administrative work. In addition, the software is getting old and hard to maintain. It is necessary to find a new system to provide integrated access to the digital library. In addition to journal delivery, the Library also has other types of materials to manage, including learning objects and Special Collections materials. Extensive research was conducted to evaluate three well-known digital library software products (Han 2004). Considering both advantages and pitfalls of the existing software and the limited functionality that this system requires, we have decided to build a tailored software package in-house using current industrial best practices. The software uses Java Servlet technologies (JavaServer Page (JSP), Java Servlet) and an open source Resource Description Framework (RDF) database called Sesame (http://www.openrdf.org/) for storage. We are currently investigating RDF for future improvement of resource discovery and cross-inferencing.

The *Rangelands* website was designed with the objective of providing an optimum customer experience. With a few clicks customers should be able to easily browse, quickly locate and get full-text articles without any need for assistance. The authors used our experience with website design and understanding of users' website navigation needs to inform our design of the *Rangelands* Archives website. In addition, we examined commercial and scholarly journal websites for guidance in designing the Rangelands site with familiar approaches in order to reduce users learning curves. The overall design follows information architecture design practices for organization, labeling and navigation. For example, the website provides consistency in global navigation and local navigation, and has appropriate mixture of graphics and text. The following figure is the front page of the website.

Figure 1: *Rangelands* website



Evaluation

To measure the usefulness of collections, there are a variety of means such as observation, surveys, interviews, and transaction log analysis. Usage measurement is one of the most important evaluation means for any digital library collection. Due to limited resources, currently we are using transaction log analysis to gain some quantitative data for usage measurement. The transaction log analysis can provide data on the volume of use (“x hits”), the type of material and users (“x users; x% users from y domains”), and related information regarding users computing environment (browsers, operating systems). There are some emerging standards for measuring electronic content, including Project COUNTER and the Association of Research Libraries E-metrics. These emerging standards were included in our systems analysis for web log analysis tools. In addition, other useful usage requirements were also included to ensure that we collect correct and necessary data. AWStats was selected as the web log analysis tool.

Future Work

Sustainability of the infrastructure and maintenance of the digital files is a constant concern. This includes object-level concerns, such as persistent identifier and ongoing digital preservation efforts to ensure permanent access to the PDF and TIFF files. There are also collection-level concerns. The project team must continue to evaluate

the appropriateness of the changing technologies and create a digital library system to host these objects.

Since the very nature of sustainability is to be ongoing, the project team is beginning to develop a long-term plan to keep this project running smoothly along with other digital Library projects. The Library will need to allocate necessary resources to keep this project operating. The project team estimates the cost to the Library will be about $3,000-5,000 per year to maintain the current system infrastructure. The cost could be lower if a generic digital library system (currently being researched) can be used in the future. Additionally, the project team will continue to pursue funding opportunities for the digitizing of additional content.

Yan Han is a Systems Librarian on the Digital Libraries and Information Systems Team at the University of Arizona Library.
Jeanne L. Pfander and **Marianne Stowell Bracke** are both Agricultural and Life Sciences librarians on the Science-Engineering Team at the University of Arizona Library.

References

Berger, Marilyn. 1999. Digitization for preservation and access: A case study. *Library Hi Tech* 17, (2): 146-151.

Fleischhauer, Carl. 1998. Digital formats for content reproductions. Available from http://memory.loc.gov/ammem/formats.html (accessed Oct 10, 2004).

Gardner, Melanie. 2004. email message to AGNIC mailing list, "[AGNIC-CC] terrific news for new projects!", April 22, 2004.

Gardner, Melanie, Jean Gilbertson, and Barbara S. Hutchinson. 2002. Partnering for improved access to agricultural information: The agriculture network information center (AgNIC) initiative. *ARL* 223: 5-10.

Han, Yan. 2005. An integrated high availability computing platform. *The Electronic Library*. (accepted)

———. 2004. Digital content management: The search for a content management system. *Library Hi Tech* 22, (4): 355-365.

Hutchinson, Barbara S., Douglas Jones, and George B. Ruyle. 2003. Building a collaborative AgNIC site as an outreach model: Rangelands of the western U.S. *Reference Librarian* 39, (82): 125-140.

Hutchinson, Barbara S., and George B. Ruyle. 2002. Partnering for better management of western rangelands: Using web technologies to get the word out. *Journal of Agricultural and Food Information* 4, (3): 75-89.

———. 2000. Wired without the barbs: Using the internet for rangeland information. *Rangelands* 22, (6): 19-22.

JSTOR. "JSTOR: The production process" http://www.jstor.org/about/process.html (accessed Oct 10, 2004).

Library of Congress. 2001 "MARC to Dublin Core crosswalk" http://www.loc.gov/marc/marc2dc.html (accessed March 15, 2005).

NISO Framework Advisory Group. 2004. A framework of guidance for building good digital collections. http://www.niso.org/framework/framework2.pdf (accessed Oct 10, 2004)

Seadle, Michael. 1998. Building on the Cornell-Yale model: Digitizing the radicalism collection at Michigan state university. *Library Hi Tech* 16, (2): 19-36.

Western States Digital Standards Group. 2003. Western states digital imaging best practices. http://www.cdpheritage.org/resource/scanning/documents/WSDIBP_v1.pdf (accessed Oct 10, 2004)

CONTACT ADDRESS

Yan Han
Systems Librarian
The University of Arizona Library
1510 E. University Blvd,
Tucson, AZ, 85721-0055
PH: (520)307-2823
Fax: (520)621-8276
Email: hany@u.library.arizona.edu