JLSC

ISSN 2162-3309 | JLSC is published by the Iowa State University Digital Press | http://jlsc-pub.org

Volume 12, 1 (2024)

Repository (R)evolution: Metadata, Interoperability, and Sustainability

Linda Eells, Julia Kelly & Shannon Farrell

Eells, L., Kelly, J. & Farrell, S. (2024). Repository (R)evolution: Metadata, Interoperability, and Sustainability. *Journal of Librarianship and Scholarly Communication*, 12(1), eP16890. https://doi.org/10.31274/jlsc.16890

This article underwent semi-anonymous peer review in accordance with JLSC's peer review policy.

ISSN 2162-3309 10.31274/jlsc.16890

PRACTICE ARTICLE

Repository (R)evolution: Metadata, Interoperability, and Sustainability

Linda Eells
University of Minnesota, Associate Librarian
Julia Kelly
University of Minnesota, Librarian
Shannon Farrell
University of Minnesota, Librarian

ABSTRACT

Introduction: Successfully managing an open-access repository requires constant attention to user community priorities in order to inform the development or selection of a platform that fulfills constantly evolving functional demands in an increasingly complex operational environment. This paper uses AgEcon Search (AES) as an example of the way that varying platforms address the metadata and other platform needs of a repository. AES is a successful subject repository with an international scope that has resided on several different platforms in its 25-year lifespan.
Elements and Considerations: Critical among the technical requirements of a repository is interoperability with other information sources and the ability to accommodate and describe different types of objects, including data. Experienced in the use of easy and widely used Dublin Core (DC), as well as Machine-Readable Cataloging 21 (MARC 21)-based repository platforms, we discuss both metadata schemas from administrative and user perspectives. Reconsidering underlying metadata issues might positively impact both technical and administrative issues that are currently restricting the development of robust, interoperable systems. As managers of AES, we are uniquely placed to discuss both technical and sustainability issues.
Conclusions: Although many institutional and subject repositories are on platforms that use DC for their metadata, other options are available. MARC, the well-established library standard, can provide the wide range of fields needed to fully and accurately describe the variety of document and data types that are included in repositories.

Keywords: Repository, Metadata schema, Interoperability

Received: 07/15/2023 Accepted: 07/04/2024

DESCRIPTION OF PROGRAM/SERVICE

Introduction

Founded in 1996, AgEcon Search (AES) is a unique, successful subject repository that has grown from 50 papers to 185,000 papers, with sustained growth of 10% per year. International in scope, the repository collects, indexes, and freely distributes open-access, full-text scholarly research in the broadly defined field of agricultural and applied economics. Content includes working papers, journal articles, conference presentations, government documents, and theses and dissertations. The contributing community consists of 350 organizations and institutions across the globe, and an average of 15,000 unique users visit per day, representing nearly every country in the world.

Platform history and requirements

AES was built by the University of Minnesota (UMN) Libraries on a homegrown platform in 1996, migrated to DSpace (DSpace 2020) version 1.4 in 2007, and migrated again to the Invenio 1.3 platform (now TIND IR in version 1.23; https://www.tind.io/ir) in 2017. As repository managers, our challenge has been to identify a platform that could offer the sophisticated administrative functionality that we require (e.g., metadata enabling accurate colocation and display of journal articles and batch uploading capability) and the user functionality that our customers demand while maximizing interoperability and sustainability for future collaborations. Also, platform requirements for an international, distributed network subject repository such as AES are uniquely complex relative to an institutional repository (IR), which creates challenges (and costs) for platform vendors and designers. A typical IR is the archive for one institution's content, with numerous submitters from within the institution uploading content to various collections. AES, in contrast, serves 350 distinctly separate institutions around the world, with submissions being made by many external contributing community members at any given time. For both contributing communities and researchers, functions such as sophisticated search algorithms; faceted results display; the ability to filter, mark, and export records; and personalization features (e.g., customized alerts) are critical to attract and retain users and thereby ensure the long-term viability of a repository. Given the need to update the AES platform to better meet all of these unique core requirements, managers initiated an investigation of possible solutions and alternate (to Dublin Core [DC]) metadata standards that are available to the repository community.

User focus

This strong focus on users is sometimes overlooked in repository planning, leading to difficulties in recruiting and retaining content and failure to thrive over the long term.

St. Jean et al. noted that "Despite the widespread recognition of the central importance of end-users to the ultimate success of an IR, we know very little about end-users" (St. Jean et al., 2011). Gonzales et al. stated that "many university faculty members are eager for an intuitive, user-friendly tool that will allow them to store, retrieve, and share their research outputs, as long as the tool is designed with their needs in mind" (Gonzales et al., 2021). Repositories must "serve scholarly communication first and foremost" because "[a]cceptance and usage by the scholarly community is crucial to sustainability" (Armbruster & Romary, 2010).

Realizing the importance of user perspectives in anticipation of a migration to a new platform, a short user survey was conducted in 2013 to assess current and desired uses of AES, which confirmed informal feedback that we had received. Users were seeking features that were not available in the DSpace instance, such as the ability to "Select multiple papers to save, export, download, or print" and "Generate a citation to a paper in a common format such as APA or MLA" (Kelly & Eells, 2013). User testing of the TIND platform (then an early version of Invenio) prior to migration would have been ideal but was not possible due to the timing and development of a new version of Invenio. The added value is that "…if a repository were designed with users' needs in mind, and took into account their behaviors and interactions with every aspect of the tool, it had the potential to increase adoption and usability far beyond numbers generally observed for university IRs" (Gonzales et al., 2021). User testing of the current version of the TIND IR is highly recommended in the future.

Although the TIND platform itself offers functionality enhancing its value, we also attribute much of the long-term success of AES to our strong relationships with, and service to, agricultural and applied economics associations and researchers around the world. These relationships are enabled by the unique pre-print culture in applied economics, with researchers sharing working and conference papers with colleagues as widely as possible for feedback prior to publication. AES provides a service that is highly valued by this community of users, and constant innovations focused on their latest needs (e.g., data management and interoperability) should ensure sustainability far into the future.

INTEROPERABILITY

It is impossible to overstate the importance of interoperability in repository platform design, but it has not been highly prioritized until relatively recently. Interoperability is defined by the Confederation of Open Access Repositories (COAR) as "the ability of systems to communicate with each other and pass information back and forth in a usable format" (Confederation of Open Access Repositories, 2011). Over the past 25 years, libraries selecting or developing platforms for managing institutional repositories, e.g., have focused mostly on creating an individual application or instance that meets local needs. Lynch defined "…a university-based

institutional repository [as] a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members" (Lynch, 2003). Ensuring interoperability meant complying with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which enables content platforms to reliably share metadata with many other platforms, including the smooth integration with local campus discovery services. The "OAI-PMH supports the dissemination of records in multiple metadata formats from a repository" although "for purposes of interoperability, repositories must [also be capable of disseminating] Dublin Core, without any qualification" (The Open Archives Initiative Protocol for Metadata Harvesting, 2015).

Interoperability extends far beyond the walls of one institution and beyond the repository and Integrated Library System (ILS) landscape. An important partner of AES, e.g., is the Research Papers in Economics (RePEc) service (http://repec.org), which indexes 2,200 content providers in over 100 countries, including all major publishers and research outlets. Inclusion of their publications in RePEc is very important to many AES community member organizations. TIND information technology (IT) managers worked intensively with AES and RePEc managers to ensure smooth and automatic daily updates of AES content in RePEc. Other external services can be readily connected with AES content due to crosswalks and export tools that are now available. Custom import or export formats can also be developed or configured, with detailed mapping enhanced by the relatively granular MARC metadata. Therefore, interoperability can be understood in a much broader context than previously considered by some IR managers.

IRs

IRs are not typically designed to meet either the archival or research needs of a broader user community outside the walls of the institution; therefore, interoperability among repositories on different platforms has not been a primary concern. Even within institutions, interoperability among various service platforms has been problematic. For example, most institutions manage their online public access catalog (OPAC) and services in an ILS, with their IR and data repository on another platform, and, often, their archives are in yet a third platform. The TIND IR capitalizes on the development of an ILS module to the CDSware platform (Silvestre, 2010) and subsequent modules added IR and digital archive functionality.

Subject repositories

Even subject repositories such as AES and arXiv (https://arxiv.org) are, by definition, designed to serve the unique needs of a specific contributor and user community (Thibodeau, 2007). A significant long-term consequence of this focus has been specialization, modification, and the

development of many silos of information that do not, in general, play well with each other. The parallel proliferation of different Descriptive Metadata schemes (or variations due to modifications) in repository platforms has resulted in unanticipated interoperability issues due to incompatibility and inconsistencies in metadata between schemes.

DC

Metadata can be shared most readily if it adheres to a common standard, and, although many standards exist and are followed, the more flexible schemes are often not consistently used. For example, DSpace, one of the most widely used IR platforms in the United States, follows the DC Metadata Element Set (DCMES; DCMI, 2012) scheme. Although this is a standard set of elements common to many repository platforms (Bankier & Gleason, 2014), the manner in which those elements are defined and populated can vary widely in each instance or institution. Often, the implementation of a platform such as DSpace requires that DC metadata elements (https://www.dublincore.org/), which are limited in scope, be modified by adding nonstandard elements to fit specific repository community requirements. Each modification introduces compatibility issues with other platforms, as well as complexities that must be replicated with each migration, e.g., to a new version of the platform. Interoperability issues arise as well, e.g., in mapping from DC to something more granular such as Metadata Object Description Schema (MODS) or DataCite.

PLATFORM REVIEW AND SELECTION

AES in DSpace

Although modifications are necessary to maximize the utility of DSpace in a specific instance, they can result in serious difficulties, particularly in moving to new platforms or even new versions of the same platform. For example, when it was time to migrate AES to a new version of the DSpace platform in 2011, we realized that all of the modifications made in 2007 (detailed in the Metadata and Functionality/AES Metadata in DSpace section) would have to be reconfigured in the new version. This process would be time-intensive and costly, and local programming restraints limited the number of modifications that could be implemented. UMN Libraries management initially recommended that we move ahead with migration to a new version of DSpace in spite of a potential loss in functionality for AES. This is understandable from an administrative perspective given that the University's primary IR (University Digital Conservancy) is in DSpace, and managing multiple platforms for different content typically increases IT expenses. However, given that the functionality of DSpace was already problematic for AES, and facing the probable additional erosion of functionality, we (as repository managers) decided to delay migration as we searched for a platform that better met our needs.

Islandora migration attempt

The first alternative platform option that was seriously considered was Islandora, which initially applies DC Extensible Markup Language (XML) metadata, often enhanced by MODS metadata. MODS is a bit more structured than DC and would provide the added functionality that we sought within a framework that was purported to be easier to adapt to specific instances than DSpace. However, an 18-month project to migrate AES to an Islandora installation proved to be unworkable. The migration project manager noted that "There were technical failures due to a lack of understanding of the complexity of the AgEcon requirement set, coupled with a lack of prior expertise working with Islandora. Ultimately, it was a perfect storm of lack of resources, expertise, and …in thinking that we could migrate to this more complex environment given our existing resources (staff and money)" (personal communication, March 19, 2024). In the end, we needed to identify another option.

Following the unsuccessful effort to implement a new AES system in Islandora, in May 2012, a Digital Repository Task Force (DRTF) was formed to consider how various selected options could effectively support UMN Libraries' diverse and complex repository and preservation services. Their recommendations were the basis for determining and implementing a digital repository technology strategy not only for AES, but also the University's IR, the University's Media repository, and University Archives needs. The DRTF was chartered by University Libraries' top leadership, with committee members representing the Web development/IT team, repository administration/managers, and metadata experts.

DRTF report and recommendation

The DRTF thoroughly investigated the pros and cons of several possible solutions, evaluating for costs (licensing and local maintenance/operating), metadata structure, expertise requirements, etc. Options included open-source and hosted systems: Hydra (with Fedora as repository); Digital Commons (BePress); DSpace; a local Drupal based solution; and some combination of these options. Their recommendation specifically for AES was as follows:

> "The Task Force reviewed three major concerns for AgEcon (distributed, simple ingest; the data model; and bibliographic export & harvesting) and could not associate these AgEcon functionalities within a single system. The primary difference between AgEcon and the UDC [University's IR] is the data model. Whereas the UDC's traditional IR data model remained unchanged, AgEcon's subject repository data model, especially its journal content, was ill-suited for DSpace from the beginning. Customizations to the data table in the AgEcon instance was the first fork of the DSpace core code and inhibited version updates. Likewise, the unique

> customization of the ingest process further deviated AgEcon from the upgrade process. Today [July 2013], the use of a SWORD API would alleviate this need to change the built-in ingest process, but it would not rectify the disparity between the metadata provided (MODS derived MARC21) and the data model (Qualified Dublin Core) to house it and to export it appropriately to bibliographic systems." (Digital Repositories Task Force, 2013)

Invenio/TIND

Therefore, from 2007 to 2017, AES remained in the same version of DSpace, whereas the UMN IR (University Digital Conservancy) was upgraded twice to newer versions. Serendipitously, at the Open Repositories 2014 conference in Helsinki, Finland, we discovered a new company marketing a spinoff of the "Conseil Européen pour la Recherche Nucléaire," or European Council for Nuclear Research (CERN) repository platform (with CERN's full support), called Invenio (now TIND IR). A full functional analysis was undertaken at this point to compare Invenio 1.1.3 and DSpace features.

Functional requirements comparison: DSpace 1.4 and Invenio 1.1.3

A full list of functional requirements in 2015 listed 47 items in six categories: Search and Browse; File Structure; Submission; Administrative Tasks; Design/Display; and Communication Tools. A specific preferred metadata scheme was not indicated because the functional requirements list was designed to lead to selection of the platform that best met the items listed regardless of metadata scheme employed. Interoperability was also not listed as a separate item, although specific functional requirements, including compatibility with RePEc indexing (discussed in the Interoperability section) and Google indexing, were included in the full list. A subset of important requirements, many of which could not be met in DSpace version 1.4, are listed in Table 1. Functionality responses in this table were provided by programmers at TIND and UMN Libraries.

This analysis was performed in 2015 and would no doubt have some additional requirements listed if performed today, and some formerly missing functional features are likely now available in DSpace.

Invenio/TIND recommendation

After spending months investigating details of the platform and company, as well as performing meticulous comparison of its functionality with the newest available DSpace update (version 3.x), the Assistant University Librarian for Data and Technology determined that

	Functionality Comparison 2015	Invenio/ TIND	DSpace
	Search and Browse		
7	Browse (and reverse order browse) by institution/journal, author, date, or subject category	Yes	No
9	Limit searches to particular document TYPE, such as Journal Article or Conference Paper (as tagged in the metadata)	Yes	Yes*Cannot do at top level
	File Structure		
11	Ability to harvest from and be harvested by OAI-compliant resources, RePEc in particular (http://ideas.repec.org/)	Yes	Yes
12	Ability to mark multiple records and email, export to citation managers, etc.	Yes*	No
13	Easily colocate and order journal articles by volume, issue, page numbers	Yes	No
14	Friendly to (and liked by) search engines, especially Google Scholar	Yes	Yes
	Submission		
18	Batch uploads without programmer intervention, using easily customized csv templates - TOOL?	Yes	No
	Administrative tasks		
24	Statistics include downloads rolled up with totals on all levels (e.g. paper, series, institution, AgEcon Search)	Yes	Yes
29	Ability to make minor updates to Web site without going through designer	Yes	No
33	Ability to move papers to different communities if they are submitted incorrectly without re-uploading (as a batch)	Yes	No
34	Allow for global edits	Yes	No
	Design/Display		
36	Automatic alerts for new submissions, refinable by subject area, by institution, etc. - via RSS feed or other mechanism	Yes	Yes
38	Page for each institution, with address, links to their Web site, links to browsing their documents	Yes	Yes
42	In search results, ability to move to next and previous records	Yes	No
46	MY AgEcon Search, with customized lists of resources (e.g. baskets, Personalization features)	Yes	No

Table 1. 2015 functional requirements (selected): Invenio/TIND version 1.1.3 and DSpace version 1.4

the perceived advantages were worth the risk of moving to a new company (Butler, 2015). Butler's recommendation stated that the "TIND-managed *Invenio* software application for supporting *AgEcon Search*, was assessed as a technology platform in the following ways: Satisfaction of Functional Requirements, Libraries' Repository Architecture Requirements,

Risk–Data Security, Risk–Financial Stability, and Potential Opportunity. The assessment assumes, based on the information provided, that this solution is a completely hosted, managed application, without dependencies on local technology staff." The risk factor was mitigated by the fact that the TIND IR platform was originally based on Invenio, formerly the CERN Document Service Software (CDSware) platform that CERN developed decades ago. TIND's developers worked closely with CERN programmers, and its founding members were initially housed at and employed by CERN.

METADATA AND FUNCTIONALITY

TIND features and functionality

The migration was completed in April 2017, and the performance of the platform far exceeded our expectations. This assessment is based on managers' daily use of the resource, as well as anecdotal feedback from users comparing functionality needs and desires prior to migration with those available postmigration. For users, AES offers sophisticated searching and sorting functionality, the ability to mark and export search results in multiple formats, and new personalization options such as "baskets" and alerts. TIND programmers worked directly with Google Scholar (GS) to ensure that AES content is highly optimized in the search engine, greatly enhancing discoverability for users worldwide. Programmers also worked with RePEc index managers to implement an automated process that replaced a labor-intense, manual monthly process in place when AES was in the DSpace platform, saving countless staff hours. MARC fields corresponding to RePEc codes (Appendix 1) were added to the TIND collection records to enable an export of records each night with corresponding "pull" from RePEc to update those listings. Both AES and RePEc currently have high visibility in GS, ranking 20th and 8th, respectively, in a recent ranking of more than 4,000 repositories worldwide (Consejo Superior de Investigaciones Cientificas, 2024). A majority of AES users discover content via GS; therefore, this high visibility is crucial.

Functional features on the back end, including global editing, batch uploading, and static page management, have greatly reduced administrative time and expense. Repository managers who are not IT professionals can now perform many of the tasks that required programmer expertise and intervention in DSpace. We attribute some of this to TIND's reliance on MARC 21 as the base metadata scheme, a choice CERN made at some point years ago when they developed their original version of the platform (Pepe et al., 2005). After experiencing years of frustration with the limitations of the DC metadata scheme (used in DSpace, in our case), we have been impressed with the administrative ease and operational functionality offered by the TIND IR platform.

MARC and DC development

Over 60 years ago, it became obvious to the library community that they needed a machine-readable standard for descriptive data, and catalogers and experts in what we now call "metadata" responded by developing different schemes to describe information objects. The first scheme, MARC (Machine Readable Cataloging) standard, was developed in the 1960s as an implementation of the ISO 2709 (https://www.iso.org/standard/41319.html) and American National Standards Institute (ANSI)/National Information Standards Organization (NISO) Z39.2 (https://www.niso.org/publications/ansiniso-z392-1994-r2016) formats for information exchange. By 1999, it had become MARC 21, a family of international standards for authority, holdings, classification, and bibliographic records. However, criticism of MARC's complexity and perceived lack of flexibility led to the quest for simpler alternatives (Tennant, 2002). Some of the most well-known alternatives are DC, MODS, and, more recently, Bibliographic Framework (BIBFRAME). However, the dominance of repository platforms (e.g., DSpace, EPrints) that use DC (Jisc, 2024) has resulted in the continued reliance on this metadata scheme, even though critics note that DC "cannot express bibliographic citation information adequately for academic papers" (Arlitsch & O'Brien, 2012).

DC in practice

DC originally included 13 (later 15) optional and repeatable metadata elements, which expanded into the DC Metadata Element Set (DCMES) with three additional elements and a group of qualifiers to aid in resource discovery (DCMI, 2012). Although DCMES has been widely adopted in the repository community due to its perceived simplicity and ease of use, e.g., by DSpace and Fedora users, the flexibility and simplicity inherent in the model have also resulted in the necessity for modifications that inadvertently created incompatibility among instances of the platform. Although all of these alternatives have advantages over MARC, the lack of uniformity in the ways that they have been implemented has made metadata integration problematic throughout the information landscape, which includes library catalogs, indexes, and repositories. Importantly, many repositories rely on metadata schemes that are not easily parsed by search engine algorithms, especially journal metadata (volume, issue, page numbers, etc.). This detailed metadata is important for elevating search results or exporting search results to citation managers. This has become more problematic as technology has evolved and search engines such as GS have developed increasingly sophisticated search algorithms and functionality.

MARC in practice

This inspires a revisit to the gold standard, MARC. MARC 21 is, to this day, the most prevalent bibliographic description scheme in use in libraries in spite of movements over the years to

create improved, simplified standard schemes. MARC is the base metadata scheme underlying every online library catalog of any size in the world and continues to serve as a universally consistent, elegant, and pervasive set of standards for bibliographic description. Some of the largest indexes in the world, including Online Computer Library Center (OCLC)'s WorldCat, as well as HathiTrust, rely on MARC as their base scheme. Although the cost of creating and applying this level of metadata to digital objects has long been a concern of some repository administrators, sophisticated standard metadata has become increasingly more important for enhancing the discoverability and visibility of collections. Search engines such as GS design search algorithms and citation creation and tracking tools based on the metadata applied to digital content. We posit that the time spent in applying sophisticated descriptive MARC metadata to repository objects is well worth the effort in increased discoverability and usability.

In choosing MARC, data entry and record maintenance training is not a barrier. Nearly every library system in the world employs at least one staff member with MARC expertise (and not necessarily a programmer). As managers of AES, we are neither trained catalogers nor programmers; however, with no significant training, we have mastered, with ease, the ~20 MARC fields that are most commonly used in the repository (Appendix 1). After years of experience using both DC and MARC metadata schemes, we have determined that establishing MARC records in AES creates no extra effort over DC, and its use offers many advantages, as outlined later in this paper.

AES metadata in DSpace

The contrast between descriptive elements in the current MARC environment and DC in DSpace is striking. With DC, our developers had to create many modified fields to accommodate and correctly display the complexities of different types of content in the repository (especially journal article metadata). This scheme proved to be seriously deficient in fulfilling many functional needs for AES, which hosts journal articles in addition to other types of content for which standard descriptive DC elements do not exist, such as conference papers. In DSpace, some base DC fields had to be modified, and, in many cases, custom fields had to be created (Table 2).

These qualifiers helped describe each issue, but correctly sorting and co-locating volumes and the issues within each volume remained elusive. Also, because these qualifiers are not standard, they are not readily recognized by large search engines such as GS as it tries to interpret and extract metadata for the creation of correctly formatted citations (including for export, e.g., to EndNote or Zotero). Although this is not a critical requirement for many institutional repositories due to the informal and internal nature of much of that content, AES is a subject

Field Description	DC field	MARC field code and indicator
Series identifier (ISSN)	dc.identifier	022__a
DOI	dc.identifier	0247_a
Language	dc.language.iso	041__a
Other classification number (JEL)	dc.subject.other	084__a
Author	dc.contributor.author	100__a
Title	dc.title	245__a
Alternative title	dc.title.alternative	246__a
Publishing date	dc.date.issued	260__c
Pre-publication date	dc.date.accessioned	269__c
Number of pages	dc.format.extent	300__a
Document type	dc.type	336__a
Series title and number	dc.relation.ispartofseries	490__a
Notes	dc.description	500__a
Abstract English	dc.description abstract	520__a
Subjects	dc.subject.classification	650__a
Keywords English	dc.subject	6531_a
Secondary authors	dc.contributor.author	700__a
Journal Format Fields *in italics*		
Date of publication	*agecon.format.ispartofname*	*773__d*
Volume	*agecon.relation.ispartofvolume*	*773__j*
Issue number	*agecon.relation.ispartofnumber*	*773__k*
Page to	*agecon.format.hasEndPage*	*773__o*
Page from	*agecon.format.hasStartPage*	*773__q*
Journal name	*agecon.relation.ispartoftitle*	*773__t*
File size	created automatically	8564_s
File link	created automatically	8564_u
Link description	created automatically	8564_y
Previous file link (non-standard)	dc.identifier.uri	887__a

DOI, digital object identifier; ISSN, International Standard Serial Number; JEL, Journal of Economic Literature.

Table 2. AES repository: DC to MARC map

repository with an international research audience, and the ability of external harvesters to correctly parse and translate the metadata elements is a critical functional requirement. As noted by Arlitsch & O'Brien, "Google Scholar has difficulty indexing the contents of institutional repositories," and the authors hypothesize the reason is that most repositories use DC (Arlitsch & O'Brien, 2012).

AES ADMINISTRATIVE FUNCTIONS

In contrast, not only does MARC easily accommodate the specific field requirements of journal articles and other repository content without any modifications or "version" issues, thereby significantly reducing development time (and costs), other cost savings have surfaced. As repository managers, we have discovered that we are able to perform many administrative tasks that would have required costly (and relatively unavailable due to competing priorities) programmer intervention when we were in DSpace.

Batch upload: DSpace

The most important of two administrative tasks that we can now perform involves batch upload functions. Although batch uploading is technically possible in DSpace, the process requires administrative access to the backend and has to be performed by programmers in the UMN library system, owing, in part, to permissions concerns, as well as technical expertise requirements. The current (in 2024) top University Libraries repository administrator notes that "Batch ingest without programmer intervention continues to be a challenge for most repository managers of DSpace installations. I don't believe there would have been any possibility to have done it on version 1.4. [B]atch importing without programmer intervention requires top administrator privileges along with a technical understanding of Simple Archive Format plus familiarity with desktop Java programming. Even then, it was somewhat unstable and would occasionally introduce errors into the database, then requiring programmer intervention." (personal communication, March 19, 2024).

The AES instance in DSpace version 1.4 also demonstrated frequent instability because the system was unable to handle more than two simultaneous submissions, let alone the hundreds of simultaneous submissions that occurred several times a year, near conference paper submission deadlines. Concerns about that instability were, as mentioned earlier, another reason to limit the use of batch upload to the Web development team. In any case, UMN Libraries' IT staff's capacity to perform batch uploads was extremely limited due to competing priorities and limited time to devote to AES. Owing to these limitations, we were never able to do a beta test of this functionality in DSpace.

Batch upload: TIND

In contrast, the availability and ease of use of the batch uploading process in TIND IR has saved AES managers, our student employees, and our user community hundreds of hours of time (and, therefore, money). We are now able to not only maintain the repository but to demonstrate continued growth of 10% per year, with zero programmer intervention. We

have used the batch upload process to add large collections from many organizations, some of whom had been requesting it for years prior to our migration to TIND. We continue to gain new customers and form new relationships weekly due to the availability and ease of the batch process. In 2021 alone, we uploaded 10,000 papers via this process, which translated to 1,000 hours or $12,000 worth of student time (students can manually upload an average of 10 documents an hour and earned $12 per hour in 2021. Effective September 2022, student employees earn a minimum of $15 per hour due to COVID-19 pandemic labor changes).

The batch upload process has also transformed the nature of the work that our student employees perform, greatly increasing the number of projects that we are able to complete in a given time period. A significant component of our operation is managing digitization projects. Spreadsheet inventories of print materials are developed and used to obtain price quotes for digitizing and then are used as the final batch upload spreadsheet. This completely negates the need for uploading the documents one at a time, as students have done in the past. This saves hundreds of hours of their time while enabling repository managers to upload thousands of documents in minutes.

Batch edit: TIND

Batch editing is another function that has saved many hours of administrative time and was not available (at least not to us as managers) in DSpace. This editing process has numerous practical applications that greatly improve the quality and consistency of the repository metadata. For example, immediately after we migrated to the TIND IR, we discovered significant instances of corrupt and incorrect data throughout the repository. Inconsistencies were easily introduced when collections were established; however, they were completely hidden from us in DSpace due to the structure of the metadata. Numerous journal titles had multiple spelling variations; volumes and issues even within a specific journal were listed in various ways (e.g., Volume 01, Volume 2); etc. Unlike DSpace, the TIND IR displays filtered search results for our journals (including volume and issue), and these inconsistencies were immediately visible and created serious filtering and sorting problems. Fortunately, via the batch edit tool, we can query a specific collection and MARC field (e.g., 773__k for issue number), view the metadata of records matching the query, and correct all erroneous records with a few keystrokes. Following migration, we spent many hours making post-migration data corrections, and the repository now has clean, consistent, and accurate descriptive metadata.

As another example, we have, on many occasions, needed to move an item, or, in some cases, many items, from one collection to another. An example might be an editor accidentally uploading a set of articles to the wrong journal issue number. In DSpace, we would have had to duplicate upload each individual article to the correct issue number and delete the incorrect article. No mechanism existed that we were aware of or had access to for correcting the metadata

for each article retrospectively, let alone correcting or moving many records with one operation. In the TIND IR, we can use the Batch Editor to identify all of the records with the incorrect information in the appropriate MARC tag (e.g., 773__k for issue number, or 980__a for collection) and correct the metadata for those documents all at the same time. We use this function weekly because, with so many external contributors and our student employees uploading documents, errors happen. Significant programmer time is saved now that we have the administrative ability to fix these errors ourselves, owing in part to the structure and simplicity of the metadata.

It is important to note that the DSpace instance of AES was a locally hosted solution, whereas AES in TIND is an externally hosted Software-as-a-service (SaaS) solution. Therefore, this analysis is not a direct comparison of administrative tasks that are possible in each. However, our experience with TIND is that the platform was developed with a focus on producing special tools to allow non-programmers to easily perform tasks normally restricted to programmers in the DSpace environment, e.g., batch uploads, static page management, user interface/display customization, etc.

SUSTAINABILITY

Distributed network model

Sustainability is always an issue for repository managers both in terms of content growth and in terms of maintenance cost. This is especially true of repositories such as AES, whereby all content is full-text and free of charge to anyone, anywhere. Sustainability in terms of content is not a problem because our growth rate has consistently been 10% per year over the last decade, but overhead is increasingly an issue. Operationally, we follow a relatively unique "distributed network" model that relies on the active participation of our member communities. Contributing organizations are responsible for content review and quality, and they either upload their papers or complete batch template spreadsheets with the requisite metadata for their collection additions. This model saves management time but continues to entail some degree of retrospective metadata correction when significant inconsistencies are identified (e.g., names entered in Given Name_Family Name order instead of Family Name, Given Name order). In TIND, submission form restrictions, auto-complete fields, limited dropdown lists, and formatting requirements have reduced the number of errors introduced by users, and managers are better able to visualize (via filters) and correct (via batch editing) errors.

Maintenance cost

Organizations are charged very little or nothing to contribute their content, the latter an option for those who take full responsibility for uploading their own content. This cooperative

model has, from the beginning, enabled continued growth with very low overhead (administrative, maintenance, and IT) costs, especially relative to most repositories, whether IR or subject. As the repository has grown, the amount of time that the two managers (currently 1 full-time employee) devote to the project has also increased. However, we have realized significant balancing reductions in costs since migrating to the new TIND IR platform in April 2017. As outlined earlier, these savings are largely due to the new uploading and editing processes that are made possible, at least in part, by TIND's platform design and use of MARC as the base descriptive metadata scheme. Savings are substantial, including much less required direct IT support (none, locally) reductions in managers' administrative time, and greatly enhanced and effective use of student staff time. Our experience has confirmed Bankier's comment that although "initially it was believed that repositories had to be open source and locally installed…a hosted service arguably has a lower total cost of ownership and is less time-consuming than running an IR locally" (Bankier & Gleason, 2014). Major IT support savings have been realized due to our movement from a locally managed DSpace repository instance to the selection of the hosted version of the TIND IR (although institutions retain the option of hosting an open-source instance of that platform locally). Although the actual cost of the TIND software and service has increased over time, even the current higher cost in 2024 is only 65% of the estimated cost of maintaining the locally hosted DSpace instance of AES a decade ago, in 2014. A repository manager recently noted that "the true cost of open-source software – [is] more like a free puppy - free to acquire but the total cost of ownership can be a bit more than often anticipated or planned" (personal communication, March 19, 2024).

NEXT STEPS

Data

Although we did not include data management as a top level or out-of-the-box functional requirement when seeking a new platform for AES a decade ago, it was definitely on our radar as a "need to have someday" function. In the current repository environment, the ability to store, describe, make accessible, and maintain data is a "highly desirable" or, in some cases, "required" functional requirement. Recognizing the importance of this issue, managers plan to obtain a grant in the near future to investigate the data archiving needs of AES users and to assess the possible platform solutions for optimally meeting those needs.

Related to this issue, the more service requirements that one platform can meet (institutional repository, digital archives, integrated library system, data management), the more likely that it can be a one-stop-shop within a library system. An important component of the TIND suite of products, in addition to TIND's IR and ILS, is their Research Data Management platform

(RDM). The TIND RDM uses a custom MARC map to DataCite elements, thereby creating a standardized interoperable metadata scheme for handling research data. This default scheme is based on the Data Cite Metadata Scheme but is fully customizable to meet more complex needs and is compliant with Findable, Accessible, Interoperable, and Reusable (FAIR) principles (https://www.tind.io/rdm). TIND's genesis as a spinoff of CERN means that their ability to handle data reflects the deep historical development of data-management technology at CERN. Data sets large and small support research papers and journal articles, with this data increasingly being required by US agencies (Nelson, 2022) and international organizations (European Commission, n.d.).

Platform assessment and improvements

Data-management functionality is a critical need that is already being assessed, but no repository is a static resource, and the scholarly communication landscape can change quickly. Identifying and responding to those changes requires continual attention to developments while adhering to criteria articulated by the international repository community. For example, AES is a member of OpenDOAR, the Directory of Open Access Repositories (Jisc, 2024), a "quality-assured, global directory" that reviews member directories prior to listing them, to ensure they meet specific inclusion criteria. AES also meets all of the "Essential" and most of the "Desired" characteristics articulated in the COAR Community Framework for Best Practices in Repositories (Confederation of Open Access Repositories, 2020).

As AES managers look to the future, actively and continually assessing the needs of the contributing community and researchers who value the repository is paramount to its long-term success. Examples of improvements implemented since the initial migration to TIND in 2017 include the automatic assignation of digital object identifiers (DOIs) to uploads, the addition of static cover pages to each download, and the capacity to add author Open Researcher and Contributor IDentification (ORCIDs) to records (albeit administratively to the backend). To enhance interoperability, TIND has developed crosswalks from MARC to other formats such as the following:

- Schema.org
- DataCite
- DC
- MODS
- Formats used by citation tools, e.g., RIS, BibTex, Zotero

Adding an author identifier field to the regular submission form, improving the cover page to capture dynamic information, and streamlining backend processes (such as the method for creating new collections) are a few improvements that would be helpful. It has been 10 years since the initial AES functionality list was created and the comparison between DSpace and Invenio was performed. An important requisite next step is the initiation of a new functional requirements list based on the current repository environment and options, including contributing community and researcher needs.

CONCLUSION

Based on our experience with both easy and popular DC- and MARC-based platforms, we have discovered many advantages to using the latter, particularly from an administrative perspective. In this fast-moving information environment with silos continuing to proliferate, adherence to the international and widely used MARC format should not be counted as a negative factor when evaluating platforms, especially when weighed with significantly improved functionality for batch loading and data management. Our experience demonstrates that the MARC format need not be a hindrance to moving forward into an environment of greater interoperability and flexibility in managing objects of all types. The TIND IR platform is an example of the flexibility and interoperability possible not only within an institution (e.g., IR, ILS, data repository, and archive) but among organizations and the repositories that they manage.

REFERENCES

Arlitsch, K., & O'Brien, P. S. 2012. Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech 30*(1): 60–81. https://doi.org/10.1108/07378831211213210

Armbruster, C., & Romary, L. 2010. Comparing repository types: Challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. *International Journal of Digital Library Systems 1*(4): 61–73. https://doi.org/10.4018/jdls.2010100104

Bankier, J. G., & Gleason, K. 2014. Institutional repository software comparison. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000227115

Butler, J. 2015. AgEcon Search: Recommendation on platform change request. Unpublished. Minneapolis, MN: University of Minnesota Libraries.

Confederation of Open Access Repositories. 2011. The case for interoperability for open access repositories. Version 1.0. https://www.coar-repositories.org/files/A-Case-for-Interoperability-Final-Version.pdf

Confederation of Open Access Repositories. (2020, October 8). COAR community framework for best practices in repositories. Zenodo. https://doi.org/10.5281/zenodo.4110829

Consejo Superior de Investigaciones Cientificas (CSIC). (2024, March). Transparent ranking: All repositories by Google Scholar. Ranking of Web Repositories. Accessed June 10, 2024. https://repositories.webometrics.info/en/transparent

DCMI. (2012, June 14). Dublin Core Metadata Element Set, Version 1.1: Reference Description. Dublin Core. https://www.dublincore.org/specifications/dublin-core/dces/

Digital Repositories Task Force. 2013. Digital repository task force findings. Unpublished. Minneapolis, MN: University of Minnesota Libraries.

DSpace. 2020. "About DSpace." About DSpace. https://duraspace.org/dspace/about/

European Commission. n.d. Open data, software and code guidelines. *Open Research Europe*. Accessed February 27, 2023. https://open-research-europe.ec.europa.eu/for-authors/data-guidelines

Gonzales, S., Carson, M. B., Viger, G., O'Keefe, L., Allen, N. B., Ferrie, J. P., & Holmes, K. 2021. User testing with microinteractions: Enhancing a next generation repository. *Information Technology and Libraries 40*(1). https://doi.org/10.6017/ital.v40i1.12341

Jisc. 2024. OpenDOAR About. OpenDOAR. Accessed March 20, 2024. https://v2.sherpa.ac.uk/opendoar/about.html

Jisc. 2024. OpenDOAR Statistics. OpenDOAR. Accessed March 20, 2024. https://v2.sherpa.ac.uk/view/repository_visualisations/1.html.

Kelly, J. A., & Eells, L. 2013. AgEcon Search user survey 2013. https://doi.org/10.22004/AG.ECON.162354

Lynch, C. A. 2003. Institutional repositories: Essential infrastructure for scholarship in the digital age. *Portal: Libraries and the Academy 3*(2): 327–336. https://doi.org/10.1353/pla.2003.0039

Nelson, A., and Executive Office of the President, Office of Science and Technology Policy. (2022, August 25). *Memorandum for the Heads of Executive Departments and Agencies* [Memorandum]. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

Pepe, A., Vesely, M., Robinson, N., Le Meur, J.-Y., Gracco, M., Baron, T., & Simko, T. 2005. CERN document server software: The integrated digital library. Accessed June 29, 2024. https://cds.cern.ch/record/853565?ln=en

Silvestre, J.J.R. 2010. *An integrated library system on the CERN document server*. [Master's Thesis, Évora University]. https://cds.cern.ch/record/1294486/files/CERN-THESIS-2010-115.pdf

St. Jean, B., Rieh, S. Y., Yakel, E., & Markey, K. 2011. Unheard voices: Institutional repository end-users. *College & Research Libraries 72*(1): 21–42. https://doi.org/10.5860/crl-71r1

Tennant, R. 2002. MARC must die. *Library Journal 127*(17): 26–27.

The Open Archives Initiative Protocol for Metadata Harvesting. (2015, January 8). Accessed March 17, 2024. https://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces

Thibodeau, K. 2007. If you build it, will it fly? Criteria for success in a digital repository. *Journal of Digital Information 8*(2). https://jodi-ojs-tdl.tdl.org/jodi/article/view/197

APPENDIX 1
MARC fields commonly used in AgEcon Search

1. Fields with asterisks are required.
2. Some fields are automatically assigned by the system to every record (e.g., DOI).
3. Journal article field (022, 773) defaults can be set on the collection's Authority Record or entered manually at upload.

022__a	ISSN
024__a	DOI (one is assigned by system; others can be entered by submitters)
041__a*	Language (short – ISO standard e.g., en or eng for English)
084__a	JEL code (Journal of Economic Literature)
242__a	Translated Title
245__a*	Title
260__c	Date Accessioned
269__a*	Date Published
336__a*	Document Type (AES uses six: Journal Article; Conference Paper/ Presentation; Working or Discussion Paper; Report; Book/ Chapter; Thesis/ Dissertation). Each type is mapped to a specific RIS citation type to obtain the most useful RIS outputs for citation managers Zotero and EndNote.
500__a	Note
520__a	Abstract
546__a	Language, full (e.g., English) - inferred from 041__a
650__a*	Subject (AES uses a list of 30 subject terms – Appendix 2)
6531_a	Keyword
700__a*	Author
720__a	Editor
773__j*	Journal Volume number
773__k*	Journal Issue number
773__q	Journal Article start page
773__o	Journal Article end page
773__t	Journal Title
980__a	Collection Number

Additional MARC fields used for/by RePEc index (http://repec.org):

190__a	Collection Name (searchable, lowest level of hierarchy)
191__a	Parent Collection/Institution Name
192__a	Provider Name (usually the same as the Parent Institution Name)

192__b	Provider homepage URL
193__a	Maintainer Name (always AES)
193__b	Maintainer Email (always aesearch@umn.edu)
194__a	Collection Type (ReDIF-Paper, ReDIF-Article, ReDIF-Book, or ReDIF-Chapter)
971__a	Collection Handle (6-character code unique to each collection, assigned by AES)
972__a	RePEc edi (7-character code unique to each Parent Institution, assigned by RePEc)

APPENDIX 2
AgEcon Search Subjects* (30)

At least one Subject is required for every paper.

*This subject list was developed in 1995 and was based on a list of 20 subjects used by the Agricultural and Applied Economics Association (AAEA)

Agribusiness

Agricultural and Food Policy

Agricultural Finance

Community/Rural/Urban Development

Consumer/Household Economics

Crop Production/Industries

Demand and Price Analysis

Environmental Economics and Policy

Farm Management

Financial Economics

Food Consumption/Nutrition/Food Safety

Food Security and Poverty

Health Economics and Policy

Industrial Organization

Institutional and Behavioral Economics

International Development

International Relations/Trade

Labor and Human Capital

Land Economics/Use

Livestock Production/Industries

Marketing

Political Economy

Production Economics

Productivity Analysis

Public Economics

Research and Development/Tech Change/Emerging Technologies

Research Methods/Statistical Methods

Resource/Energy Economics and Policy

Risk and Uncertainty

Teaching/Communication/Extension/Profession