JOURNAL OF ARCHIVAL ORGANIZATION
2015, VOL. 12, NOS. 3–4, 198–215
http://dx.doi.org/10.1080/15332748.2015.1150104

Routledge
Taylor & Francis Group

Legacy Matters: Describing Subject-Based Digital Historical Collections

Christian James[a] and Ricardo L. Punzalan[b]

[a]National Agricultural Library, Beltsville, Maryland, USA; [b]University of Maryland College of Information Studies, College Park, Maryland, USA

ABSTRACT

Evolving institutional structures and missions affect the metadata and digitization efforts of a cultural heritage institution. This article discusses the institutional challenges the National Agricultural Library (NAL) faced as it gathered contemporary and historical federal dietary guidance publications into a digital collection. The Library has over time used a variety of descriptive subject and classification schemes as well as a variety of encoding mechanisms, each in response to the administrative as well as technological changes and challenges in the repository. As the Library began compiling content for its Historical Dietary Guidance Digital Collection (HDGDC), it confronted an array of records dispersed across various series and collections with heterogeneous metadata, which are legacies of centuries-long institutional evolution. The authors consider the implications for archives and special collections in creating subject-based digital collections from items dispersed across institutional holdings.

KEYWORDS

legacy data; descriptive metadata; subject-based digital collections; descriptive data re-use; controlled vocabularies; historical dietary guidance; U.S. National Agricultural Library

The National Agricultural Library (NAL), a United States national library and a leading collector of agricultural and related information worldwide, has launched a digital collection of historical dietary guidance publications from the early 1900s to the present. This effort required library staff to examine century-old catalog records and inconsistent keywords and weed out records unrelated to the collection's theme. The challenges NAL faced in merging dietary guidance publications are related to its historical efforts of collecting and organizing these materials. Organizational changes across the U.S. Department of Agriculture (USDA) and other federal agencies have caused steady shifts in the library's mission and collecting activities. The contents of the Library's collections and their descriptions are the results of evolving missions, cataloging systems, and collecting practices. These, in turn, directly affect successive library initiatives to represent its historical collections.

This article will examine the development of the Historical Dietary Guidance Digital Collection (HDGDC), which is a compilation of federal government publications. This collection is a component of the Fedora-based National

CONTACT Ricardo L. Punzalan punzalan@umd.edu University of Maryland College of Information Studies, 4117-J Hornbake Building, South Wing, College Park, MD 20742.

Agricultural Library Digital Collections (NALDC) repository and content-delivery system, which features historical USDA publications, peer-reviewed journal articles, and other agriculture-related documents. HDGDC allows users to browse and perform faceted searches through over 1,000 keyword-indexed digitized and born-digital documents in OCR-enabled PDF format. It also integrates these documents with other NALDC collections related to food, botany, entomology, and other agricultural subfields. Health professionals, historians, and other audiences may view or search through this content to research and reference past ideals of food selection and diet composition.

Insights in this article came from a yearlong digital curation fellowship between NAL and the University of Maryland's College of Information Studies. During this period, the authors were involved in further developing the HDGC by defining the scope of the collection and in identifying content from the Library's vast general collections through research and data mining. Thus, this study benefits from direct interactions with key personnel and close involvement in the planning and design phase of the HDGDC.

We begin this article by examining available definitions of legacy systems and legacy data. We then provide a literature review looking into descriptive data and its transformation into successive description systems. Next, we delve into the details of the case, highlighting the development of legacy collections, legacy data, and metadata systems at NAL. We conclude with a discussion of the implications of this case for archives and special collections. We find that institutions' past collecting, description, and encoding processes both enable and complicate efforts to compile items in new contexts.

Since its creation in 1862, NAL has responded to new policies, needs, and capabilities by adopting new classification schemes, description and encoding standards, and digital content-delivery systems. The Library's digital systems have evolved from a citation-based online catalog to multiple content-delivery systems. This case study examines the limiting effects of three changing practices, systems and standards: (a) heterogeneous subject terms; (b) classification systems, and (c) description formats. Each of these has limited the Library's ability to efficiently compile content for digital initiatives.

Our study demonstrates the complexity of combining digital sources into a single new collection. Digital technology allows libraries and archives to represent digital surrogates in new online contexts. Inconsistent metadata accumulated through decades, or even centuries, of collecting can impede both institutions and users from effectively aggregating or using subject-based digital collections of diverse origins.

Digital projects are dependent on institutional policies, legacy systems, and legacy data, even if past efforts have failed to anticipate the needs of contemporary information professionals. This study has broad applicability beyond its setting and will make important connections to the archives and special collections fields. Our findings are especially useful for information professionals interested in compiling dispersed subject-based digital collections and re-using multiple forms of legacy data.

Before archivists and manuscript curators can take full advantage of digital technologies' capability to unite or combine content items, they should first examine institutional histories to identify conflicts in the content and form of their metadata.

Definitions

The phrase "legacy data" is frequently used in scholarly literature but rarely defined. The Society of American Archivists (SAA) glossary defines "legacy system" as a system "that is difficult to modify because the software is written in a programming language that is no longer common."[1] The cost of maintenance is a distinguishing characteristic of a legacy system. As the SAA glossary points out, the cost of maintenance "will likely be greater than the value of the information it contains or greater than the costs of upgrading the system." Nonetheless, this definition insufficiently explains legacy data as it is often understood in the library and information science community. In libraries and archives, legacy data may not necessarily be dependent on outdated, cost-heavy systems. Some legacy data, like Machine Readable Catalog (MARC) records or finding aids in Microsoft Word, are inexpensive to maintain and remain widely used, even as they are the source of more recent, alternative forms of Extensible Markup Language (XML)-based data. The Library of Congress's Encoded Archival Description (EAD) glossary comes somewhat closer to capturing the essence of legacy data by defining it as "finding aids created prior to implementation of EAD."[2] Even in archives, this context-dependent definition is too specific and neglects preliminary inventories or other finding tools. These definitions also unnecessarily focus on format or medium instead of content and semantic meaning. A new, working definition of legacy data is necessary.

In this case study, we define *legacy data* as any past descriptive information to be transformed into a new, more modern standard or system. Legacy data does not necessarily imply obsolescence or abandonment, as older data such as print finding aids may still be used in a research room concurrently with their EAD-tagged version. Nor does legacy data imply dependence on any particular system or the cost required for its maintenance. Finding aids may be in either print or electronic formats, and the cost of maintaining either could be negligible. Legacy data may also refer to content as well as its medium.

Transformation is the defining feature of legacy data. It is transformed as it is reused to become an entirely new collection of data conforming to a different descriptive standard, encoding standard, or format. We may therefore define legacy systems as any medium or encoding format that conveys or configures legacy data. However, the focus of this article is on legacy data, not legacy systems, and the problems that can arise upon migrating legacy data to new standards and systems.

A working definition of "subject-based digital collection" is also necessary to properly describe the end result of the case study at National Agricultural Library. Our working definition is approximate to, but distinct from, the definitions of "thematic digital collection" offered by Carol Palmer and John Unsworth.[3] Unsworth

emphasized, among other features, an interdisciplinary collection of digital primary sources designed to support research. For Palmer, thematic digital collections can also involve users collaborating with libraries and archives to unite items across scattered and static repositories. Neither of these definitions is ideal. They may not capture the richness or innovation of certain digital humanities collections.[4] Palmer and Unsworth's definitions also suggest close associations to academic research and scholarship, which does not match NAL's experience compiling thematic digital collections. Nonetheless, the emphasis on uniting institutions' interdisciplinary digital sources fits the subject-based digital collection at NAL described here. Our use of subject-based digital collection more closely matches that of Liz Muller's "subject-based aggregation of digital or digitized media from many collections," a distinct type of published digital collection that is "created for access reasons, frequently for a particular scholarly or popular audience."[5] Nonetheless, our use of subject-based digital collection may have implications beyond this unit of Muller's typology, such as digital exhibitions or the addition of born-digital collections to non-digital collections.

Literature review

The issue of descriptive data reuse and its transformation into successive description systems has previously been explored in archives and allied disciplines. We find that the archives, library, and digital curation fields have each recognized the need to re-use existing descriptive data to build upon successful classification and location systems and manage large volumes of records. We also recognize a need to assess the impact of evolving organizational missions, functions, and structures on the character and limitations of each institution's descriptive systems.

We position our research into a long-standing review of the application of traditional archival functions and principles to electronic environments. At least as early as 1993, archival scholars have argued that collecting institutions must evolve by altering their internal work processes and philosophies to successfully preserve electronic records.[6] In light of the unique characteristics and preservation needs of electronic media, scholars have noted the inadequacy of the traditional "archival paradigm," from the core functions of appraisal, arrangement, and description to the principles of provenance and original order.[7] Yet for every new paradigm that collecting institutions may construct, most of them must reckon with a vast, existing print collection. To bring these collections into the digital environment has required the re-use and transformation of existing data.

Archival scholars have previously discussed the importance of re-evaluating and re-using descriptive information. For instance, Margaret Hedstrom and John Leslie King demonstrated that the gradual accumulation of content, and the evolving imposition of systematic order upon it, forms society's "epistemic infrastructure."[8] Contributions to this infrastructure range from encyclopedias to national libraries and very well include classification rules and systems. The "knowledge economy,"

as exemplified by commercial information services such as Google and Amazon, is "deeply dependent on the products of traditional infrastructure" such as the Library of Congress Cataloging in Publication system (which also facilitates the dissemination of bibliographic information of soon-to-be published books to libraries, bibliographic services, and book vendors around the globe).[9] In this sense, the accumulating mass and re-use of descriptive information may be understood as a contribution to this infrastructure.

As early as 1989, David Bearman noted the utility of acquiring and repurposing indexes and finding tools to expedite the processing function of archives.[10] Anticipating the electronic storing of descriptive information, Bearman predicted that "when [archivists] attempt to control information in electronic form … what they will find may be a surprise: archival description systems have been metadata systems." The author called on archivists "to develop systems that enable descriptions of records to grow dynamically from their entire history of creation and use." Twenty-five years later, it is still unclear how descriptions evolve during this history of creation and use. By highlighting the example of the National Agricultural Library, this article will attempt to illustrate these transformations and evolutions in a specific setting.

The advent of EAD in the mid-1990s prompted additional reevaluation of existing library and archival description systems. EAD's principal architect, Daniel V. Pitti describes the encoding schema as a latest iteration of a chain of legacy systems, including the National Union Catalog of Manuscript Collections (NUCMC), MARC records and the Archival and Manuscripts Control (AMC) format of the Anglo-American Cataloging Rules (AACR2).[11] Pitti concluded,

> The economic benefits of sharing cataloging that motivated libraries were not available to archivists, whose collections are mostly unique. Nevertheless, archivists wanted to make their materials more accessible, a professional objective they shared with their library colleagues. This desire provided the motivation to explore and eventually embrace MARC AMC and APPM [the Archives, Personal Papers, and Manuscripts standard], the success of which convinced the archival community of the value and importance of encoding and descriptive content standards. Further, archivists were inspired to want to go beyond summary descriptions and to find a way to provide access to the full, detailed finding aids that constitute the heart of all efforts to make archival collections accessible.[12]

In making the new EAD, archivists relied on older standards and print finding aids; by doing so, these print finding aids now became legacy data.

The ability to migrate from paper to markup language was not only the result of technical innovation. At the dawn of EAD, scholars questioned the sanctity of the finding aid on theoretical grounds. Archivists today acknowledge that past forms of representation in libraries and archives are never complete, final versions but instead imperfect yet necessary tools in a constantly ongoing project of information storage and retrieval. Even as standardized description rules and vocabulary "lend a certain aura of objectivity" and discourage editorializing or variation, arrangement and description are inherently subjective.[13] Drawing on Bearman and the work of Terry Cook, Wendy Duff, and Verne Harris expose descriptive standards as sites for

constructing meaning and narrative, while acknowledging their utility for twenty-first-century digital endeavors.[14]

In "Archival Representations," Elizabeth Yakel made the connection between EAD and this reevaluation of legacy systems even clearer.[15] As MARC and EAD succeed analog representational systems such as card catalogs and finding aids, each system becomes a "representational artifact" that "contributes to the knowledge base of the repository at the same time it changes it."[16] This concept of contribution and transformation to the knowledge base is essential to the story of legacy data transformation. Furthermore, once archivists acknowledge that the finding aid or any other description and retrieval tool are subjective representations, these tools are no longer sacrosanct; other, newer description systems may be just as, if not more, useful or effective. By recognizing the subjectivity of past descriptive standards and systems, we can also acknowledge and study the substantive changes in description that may occur during the transition to new systems.

Armed with both technical innovation and critical theory, archivists have made changes to traditional finding aid access to address practical concerns for the user experience. A wide body of literature has also addressed the fact that finding aids are not the optimal format for users of archives to read and interpret, especially online.[17] The finding aid can and has been replaced by a variety of different online access systems. J. Gordon Daines and Cory L. Nimer found that online "single-level display" systems like Access to Memory (AtoM), which prioritizes single description components over hierarchies and front matter, are the most ideal for users of archives.[18] In a study of three unrelated online access systems, Jane Zhang found that archives have reused finding aid data to create new access points, facets, and virtual series to enhance user navigation through description.[19] In each of these new access solutions, the traditional finding aid becomes legacy data necessary for creating a new description system.

Libraries too have recently come to see legacy data and legacy systems as imperfect constructions, but must still reckon with individual histories to fully understand these legacy systems and their implementation. Library card catalogs had already been transformed into online public access catalogs (OPAC) in the 1970s and 1980s. Compared to archival descriptions, OPAC records have a higher degree of uniformity due to coordination from the Library of Congress' Cataloging in Publication program and the Online Computer Library Center (OCLC). With the advent of XML and digitization projects, libraries have become more sensitive to quality control and institutional practices on their digital initiatives. For example, Naomi Dushay and Diane I. Hillman compare transitions from structured MARC standards into XML-based metadata to the "wild west" because of a lack of institutional training and preparation, and propose automated systems to solve personnel challenges.[20]

Surveys of libraries indicate that different metadata schemas may not be interoperable in part due to a lack of documentation from individual library departments.[21] Junli Diao and Mirtha A. Hernández see the quality of metadata as both a crucial

element to digital projects and a result of a particular institutional role: the cataloging librarian.[22] As revealing as each of these analyses are, each are rooted in a static, present-minded conception of the library; no past personnel categories or evolving roles for librarians are discussed.

The re-use of legacy data is also a theme in the literature of the relatively new area of digital curation, as the Open Archival Information System (OAIS) Reference Model demands the creation of preservation metadata for content in digital repositories. Andrew Wilson, drawing upon the Australian post-custodial theory of archives, envisions the re-use of record-keepers' data in the preservation descriptive information component of OAIS Information Packages.[23] In practice, this process is difficult because records producers often have difficulty creating reliable metadata. Although this can create "metadata friction" and impede interoperability, it may be alleviated through institutional measures ranging from conversations to annotations and ad hoc tools.[24] This study reinforces the findings of our archival and library science literature review that institutional practices affect metadata quality and its potential for re-use. Far from being limited to this relatively new field of digital curation, this relationship exists in many information environments and can be documented in an institution such as the National Agricultural Library for at least a century.

In an era of exponentially increasing digital content, re-use of existing descriptions has become more practical in archival processing because it may be more efficient than creating entirely new descriptive information. To expedite the reduction of processing backlogs, archivists have proposed and attempted using existing content lists and inventories supplied by records creators.[25] Most recently, the introduction of metadata management systems such as Archon, Archivists' Toolkit, and ArchivesSpace has prompted additional evaluations of legacy representation systems. For Carrie Hintz and Cassie Schmitt, migrating to ArchivesSpace forces a reckoning with multiple legacies: past descriptive standards and practices, file structures (including the design of relational databases), and the office structure of the archive (in these cases, university archives and special collections).[26] These legacies can be significant, as Hintz cites collections at the Columbia University Rare Book and Manuscript Library that date back to the institution's origins in 1754.

The existing literature acknowledges both the importance and limitations of the content and form of legacy data. Nonetheless, most of the studies described above are theoretical in nature and do not involve detailed inquiries into the histories of institutions, their legacy data or their impact on current digital projects. Focusing on a particular case will validate the existing research in empirical ways.

This article will contribute to the body of literature by explaining the historical and institutional context of collections and their legacy systems' inception, obsolescence, and reuse. The NAL is an excellent venue for such a discussion, as its long history and evolving mission present several points of

divergence in collecting and representing activities. These points of divergence—from being a department library to a national library and later a special collections library and digital repository—became apparent in a review of the library's holdings during the creation of a subject-based digital collection. In the following sections, we will examine this collection and its institutional context.

Creating a subject-based digital collection: Institutional context

Dietary information emanates from various agencies within the U.S. federal government. This type of information includes general consumer health publications, publications addressing chronic disease prevention, and information for food program recipients. To appreciate the complexity of consolidating federal dietary information, it is important to discuss the NAL's historical role in the production of dietary information and the increasing efforts several other key agencies in developing food and health related information. A historical look at the evolution of this effort demonstrates the need to consolidate this content into a digital collection and re-use historical legacy data.

NAL's decision to create a subject-based collection of federal dietary guidance publications relates to the USDA's historical role in nutrition research. In the late nineteenth century, the USDA played a leading role in the science of modern human nutrition. Scientists have long known that different foods and ingredients produced different reactions in human and other animal bodies.[27] By the mid-twentieth century, the USDA was not the only agency involved in dietary guidance promotion. During World War II, independent offices such as the War Food Administration and the Office of Defense, Health and Welfare Services promoted nutrition-related messages. By the 1980s, health professionals became increasingly concerned about rates of diet-related chronic diseases and conditions.[28] In response, organizations such as the U.S. Department of Health and Human Services (HHS) and the Centers for Disease Control (CDC) published materials aimed at preventing disease related to poor nutrition. HHS worked with USDA to promote the enduring Dietary Guidelines for Americans series, and HHS and CDC both printed supporting publications independently. The Food and Drug Administration (FDA), responsible for mandating Nutrition Facts labels on packaged food products since 1992, also began promoting awareness of nutritional values.

The NAL initiated its effort to unite these historical federal dietary guidance publications in a digital collection as a response to frequent public and government researcher demand for the content. Important, early decisions in the project involved the location of the source print documents and the choice of digital repository. NAL staff had initially considered digitizing manuscripts and publications related to dietary guidance policy formation from within the library's Special Collections. They also considered housing these digitized documents on a stand-alone website within the nal.usda.gov domain. Other possible strategies included collaborating with other libraries to ensure the comprehensiveness of the collection. Key decisions that moved the project forward were the choices to only digitize items

from NAL's central holdings and to house these digital surrogates in the NALDC system. NAL reached these resolutions at the beginning of the 2014 Fellowship in Digital Curation with MLS candidates from the University of Maryland and allowed selection work to move forward.

As work on the project progressed, project staff encountered heterogeneous catalog metadata that complicated efforts to compile a comprehensive collection of publications with similar themes and purpose. This heterogeneity was not the result of haphazard practices or poor cataloging work. Instead, these inconsistencies were the result of shifting policies and systems. Because the collection of content spanned an entire century, various levels of organizational development and cataloging innovation had left imprints on the descriptive data that remained. As our literature review has established, institutional histories and structures affect the form and quality of descriptive data. The history of the NAL includes a series of legacy standards and systems that help us understand this impact in practice. This history demonstrates the complexities of aggregating and integrating dispersed items into a subject-based digital collection at any library or archive.

Progression of national agricultural library legacy collections and systems

Legacy systems and legacy data at the National Agricultural Library are all part of its history dating back to the mid-nineteenth century. The creation of a new cabinet-level agency devoted to agriculture was the result of lobbying from the U.S. Congress, state agencies and agricultural societies concerned about protecting farmers' interests. An act of Congress established the USDA in 1862 to create, acquire, preserve and disseminate scientific knowledge related to agriculture.[29] One provision of the legislation required the establishment of a Department library.

The USDA Library managed heterogeneous collections early in its history. The initial collection of the Department library itself was a massive acquisition of legacy content. The U.S. Patent and Trade Office possessed a wealth of information related to agricultural patents, approximating about 1,000 books and journals. The library received appropriations to purchase new acquisitions, and early collection areas included botany and entomology. The Library developed sophisticated location systems, adopting standardized catalog cards and a dictionary catalog system in the late nineteenth century. As the USDA grew, it began to author a significant number of its own publications. In 1899, Chief librarian William P. Cutter (nephew of famed librarian Charles A. Cutter) identified demand among borrowing libraries for Department publications.[30] As a result, the library spent significant effort and expenses collecting these publications and creating and distributing card catalog records for them.[31]

The USDA experienced frequent restructuring during the mid-twentieth century. Various individual bureaus were created under the control of the central Department. Each of these Bureaus possessed their own libraries and collections. For example, the Bureau of Home Economics possessed its own library distinct and physically separate from the larger, central Department of Agriculture Library. In 1942,

the Department Library acquired and consolidated all of these individual libraries. This large-scale acquisition recalls the initial Library collection of Patent and Trade Office documents. Once again, the library integrated into its core collection an entire collection of materials acquired and indexed under different directions and circumstances than its own.

In 1962, the USDA library became the National Agricultural Library (NAL) and joined a network of National Libraries, including the Library of Congress and National Library of Medicine. This new institutional mission affected collecting policies in two major ways. First, although NAL retained reference and research duties for the USDA, it now acquired copyright deposit items from the Library of Congress in its specialty collection development areas. Second, the Library was also later given the authority to acquire manuscript collections. These developments would have important impacts on the collecting process for the dietary guidance digital collection.

Although collection practices evolved, new technology led to the creation of new information systems. As libraries nationwide began converting card catalogs to electronic format, NAL created its Agriculture Online Access (AGRICOLA) database. This database was accessible in a variety of ways, including data file and CD-ROM. After the popularization of the Internet, the public could interface with library content online through a variety of unique NAL retrieval tools such as Gopher and ISIS. By the late 1990s, AGRICOLA was an online public access catalog with converted MARC records. In the 2010s, catalog records for digitized holdings were migrated to the NALDC in the Metadata Object Description Schema format. Other current access systems include the Internet Archive, with whom NAL has collaborated to make multiple library collections available on archive.org with the Dublin Core metadata schema. Besides these access systems for monographs, NAL currently maintains an Agricultural Data Commons digital repository, which hosts open statistical research datasets and database files, as well as a Digitop interface for subscription journal articles. The Special Collections unit now maintains digital collections in Omeka, rendering archival descriptions in the Dublin Core schema. NAL staff still envisions upcoming migrations to new access and retrieval systems in the future to consolidate content or accommodate emerging demands. This continuing accumulation of descriptive and access systems influences, and will continue to affect, digital collecting efforts.

Impact of legacy data on digital collection efforts

Legacy descriptive metadata is vital for digital work and both enables and limits new digital collections. The impact of institutional history and legacy descriptive systems was apparent throughout the compilation of the HDGDC. This section examines the ways that three evolving library tools—heterogeneous subject terms, classification systems, and descriptive systems—each limited the Library's ability to compile a thematic digital collection.

Heterogeneous subject terms

NAL catalog entries contained no metadata tags, or simple combination of tags, that would instantly produce a list of federal dietary guidance publications. Although dietary guidance publications have existed for a century, the category of "dietary guidance" itself is subjective. A recent USDA document precisely defines the term as, among other things, "recommendations on intakes of nutrients and other food components, suggestions on what to eat (food guides, sample menus, and recipes), and how to buy, store, and prepare foods to achieve good nutrition."[32] This effective, contemporary policy-based definition is too complex for cataloging purposes. In fact, the definition demonstrates how the category actually encompasses a wide array of specific topics, including nutrients, food selection, and cooking. The variety of possible subject terms and classifications disperses these publications across a library's holdings.

This concept of "dietary guidance" as a dispersed collection of subject terms is underscored by the identification of publications in series. Although many dietary guidance pamphlets had been printed in series, these series also contained serial issuances that were not related to food consumption in any way. For example, several landmark food guidance publications in the early 1900s were published in USDA's *Farmer's Bulletin* series. This series included topics such as pest control and crop preservation; food selection and preparation were, in this context, just one concern among many related to the lifestyle of rural families. In another long-running series, *Homemaker's Chat*, food selection and preparation advice was one of many subtopics appearing in the field of home economics. Federal agencies had yet to consider food guidance as a particular policy issue to single out for its own administration. Unifying these dispersed "dietary guidance" publications becomes even more challenging when analyzing legacy data.

Regardless of the subjectivity of "dietary guidance" as a category, certain controlled vocabularies exist for locating relevant publications based on subject descriptions. However, a variety of vocabularies populate NAL's current and legacy systems, introducing a problem of authority control. For nearly a century, the Library cataloged content with its own National Agricultural Library subject headings. The current catalog, AGRICOLA, identifies content with Library of Congress Subject Headings (LCSH); for legacy content, LCSH was retroactively applied (See Table 1). For example, the 1924 publication "Rice as Food," which boasts the nutritive value of rice, was originally cataloged with the NAL subject term, "Cookery (Rice)," but currently has the LCSH-compliant subject heading of "Cooking (Rice)" in accordance with Library of Congress authority records. Although the migration to LCSH and subsequent authority control should make for a seamless integration of keyword browsing and searching in AGRICOLA, legacy terms still appear in legacy systems such as the dictionary and card catalogs. Furthermore, these changes in subject terms demonstrate the subjective nature of these categories.

In some less frequent instances, neither the legacy data nor or its current version provides any descriptive metadata at all. Certain AGRICOLA catalog records,

Table 1. Varying Subject Terms Across Legacy Systems.

Catalog entry	USDA library subject authority term	LC subject heading in AGRICOLA	Subject category code	Digital surrogate subject terms
Rice as Food (1921)	Cookery (rice)	Cooking (rice)	Q504; T300; U000	Internet Archive: Cooking (rice)
Vitamins A, B, and C in Foods (1926)	No subject terms on card catalog record	[blank]	[none available]	[no surrogate available]
Food for families with school children (1948)	No subject terms on card catalog record	Food; nutrition; children—nutrition; families—economic aspects	[none available]	Internet archive: food; nutrition; children; families
Food guide for older folks (1974)	No subject terms on card catalog record	Older people—nutrition	U600	NALDC: [blank]

Note. AGRICOLA = Agriculture Online Access; NALDC = National Agricultural Library Digital Collections.

including many federal dietary guidance publications, have no subject terms. This may be due to a lack of descriptive data created at the initial cataloging process or a loss in translation during the transition from legacy systems. For example, the 1974 edition of the recurring title "Food Guide for Older Folks" features the LCSH term "older people—nutrition" in AGRICOLA but no subject terms at all in its NALDC record. The AGRICOLA record for "vitamins A, B, and C in foods" contains no NAL, LCSH or other keywords of any kind. Similarly, a scattered number of records possess subject category codes, an alphanumeric string unique to NAL designating over 200 topics. This inconsistency makes comprehensive catalog searches difficult for both library staff and users alike. A few categories, including T100 ("Nutrition and Health Education"), were useful for identifying content of interest. Without a consistent application of the codes across the entire catalog, this usefulness is limited. From a practical perspective, this is a problem of authority control that could be addressed by a vendor. Nonetheless, the heterogeneity of subject terms and codes, and the limits on their availability, considerably complicates the research and compilation of records on a theme.

Because of another unique NAL mission—cataloguing the work of USDA scientists—the application of heterogenous descriptive data continues today. One unique collection in the NALDC is a compilation of journal articles authored by USDA researchers. These are indexed using a unique, discipline-centric subject authority, the National Agricultural Library Thesaurus (NALT). (Further demonstrating the inherently arbitrary nature of descriptive metadata, NALT terms are periodically revised; the 2014 NALT Thesaurus is the Thirteenth Edition). In the digital collections, these journal articles are cataloged at the same bibliographic level as the items in the Historical Dietary Guidance Digital Collections, even as their subjects are described in different vocabularies. While journal articles lie outside the scope of this particular thematic digital collection, it is nonetheless likely that nutrition-related research appears within the journal articles and could be of use

to users of the dietary guidance collection. The heterogeneity of subject terms—the result of competing institutional missions—limits the ability of library patrons and staff to effectively navigate between different types among the repository's rich, diverse content.

Heterogenous classification systems

While compiling a century of dietary guidance publications, the problematic legacy description data is not limited to subject terms and codes but also includes classifications. The original NAL classification system was primarily numeric, starting with a number code that matched the determined top subject level. It was only in the 1960s, as the USDA Library became a National Library, with new ties to the Library of Congress, that NAL began to catalog each item with the Library of Congress Classification system. Older collections were not re-classified to Library of Congress Classification. Complicating this further is the fact that some of NAL's earliest government publications were classified not by subject but by government agency, with a letter forming the first part of the classification string instead of a number. As a consequence, project staff compiling the historical dietary guidance collection could not simply browse stack areas to effectively identify content based on subject; various subjects could be located in at least three different places. Neither could they browse by legacy call numbers in the AGRICOLA catalog, as the system is not built to sort documents according to this scheme. (This problem is common among government document collections within university libraries.)[33] Like the legacy subject terms, these varieties of classification data were the direct result of administrative changes throughout NAL's history.

Heterogenous descriptive systems

The change to a National Library incorporated yet another description method to the USDA Library: the finding aid. With the legislative authority to collect manuscript collections, the library now added hundreds of research collections. As with many archives and special collections, these collections are described in a variety of ways, from a full, *Describing Archives—A Content Standard* compliant finding aid to simple container lists; some have little or no description or finding tools whatsoever. Several collections, including personal papers of retired USDA employees, are related to USDA nutrition research and promotion and include monographs that could be included in the Historical Dietary Guidance Digital Collection. However, NAL manuscript collections are described at the collection level in AGRICOLA and do not indicate the title or other bibliographic data of the monographs within each collection. Without researching each nutrition-related collection box by box, it is impossible to identify the relevant content without intensive archival research. In this situation again, the administrative decision to acquire manuscript collections limited the type of descriptive data available while compiling a digital collection.

Summary

Because of the many challenges resulting from heterogenous subject terms, classification systems, and description systems, project staff examined all NAL catalog records created by the government agencies responsible for dietary guidance over the last century. This process took considerable effort. First, it required careful research to select the correct name authorities of the authoring agencies; next, it required extensive examination since these corporate authors were responsible for over 5,000 catalog records in AGRICOLA. Project staff reviewed each record line by line and eliminated 80% of the records that did not appear to meet the criteria developed by the 2012 USDA memorandum. Even after sidestepping the problem of heterogeneous subject and classification data, project staff still faced another impact of data migration. Reviewing titles of historically significant dietary guidance publications mentioned in contemporary secondary literature revealed at least two publications, "National Wartime Nutrition Guide," and "Planning Diets by the New Yardstick of Good Nutrition," that were present in the Dictionary Catalog but not in the online AGRICOLA catalog. These records had likely failed to migrate from the Dictionary Catalog system to the online public access catalog. Successful selection of the project content required that the migration of legacy data did not circumscribe the content available for selection.

Administrative and technological changes in the National Agricultural Library's history shaped the data that describes its collections. As the USDA Library started to amass specialized research literature, it used a unique subject term and classification system to precisely represent its holdings. As it began to collect more government agency publications, it used a modified classification system along with a refined and thorough card and dictionary catalog retrieval mechanism. Upon becoming a national library that acquired items through copyright deposit, it converted to Library of Congress classification, and later, migrated legacy subject data to Library of Congress Subject Headings. As it converted to an online public access catalog and a digital repository, the Library encoded this legacy data in new formats and schemas. Each of those decisions impacted the ability to easily identify dietary guidance publications by assigning necessarily arbitrary subject terms, imposing heterogeneous classification systems, or in rare cases, erroneously removing subject data or records altogether.

We now discuss some of the implications and findings of this research and their importance in archival contexts.

Implications for archives and special collections

This case study has important implications for archives and special collections by highlighting problems that may occur as a result of metadata migration. In archives and special collections, the migration of descriptive metadata has found standardization in EAD crosswalks. Applications such as ArchivesSpace, which is open source and managed and governed by many of its users, promote standardization

and interoperability across institutions. However, conversion to EAD is only possible for fully processed collections, and using various conversion mechanisms to import preliminary descriptions into a system such as ArchivesSpace could result in the same heterogeneity depicted in this study at the National Agricultural Library. This risk could be even greater when converting to various access systems such as Omeka that may not have a standardized conversion process. Archives have the capacity to achieve interoperability between processed and unprocessed collections but should be aware of the difficulty of perfecting it.

The backlog of unprocessed archival collections is not the only obstacle to achieving this interoperability. Processed collections themselves can be inconsistent based on heterogeneous descriptive data. These processed collections may have already been described in a variety of standards over many years of archival custody, undergoing a number of editorial changes and format conversions well before they are migrated to EAD. Archivists should consider how their processed collections and their descriptions are already the result of institutional histories and administrative decisions. The choice to process and describe a collection is itself a deliberate decision based on factors such as collection areas, donor relations, or even funding and staffing at a particular moment in time. Therefore, the matter of interoperability between processed and unprocessed collection metadata should not just be a technical question but also a historical consideration.

In this situation, the National Agricultural Library Special Collections had a potential role in contributing content to a digital collection but did not possess the descriptive data to identify records at the item level. Archivists have neither the mission nor the resources to describe at such a granular level but can anticipate demand for individual items to be reproduced in a subject-based collection. Compiling a research collection based upon a theme also reinforces and challenges the importance of provenance. Building a digital collection based upon topic creates a context equally as powerful as provenance. The HDGDC will allow researchers to more effectively research a century of federal food messaging and policy. By examining these records side by side, out of their original context, researchers can make interpretational connections that they may not have been able to make otherwise. Increasingly, researchers of archival collections expect to interact with content in this manner, as they come to view databases as a form of cultural expression, and simultaneously interpret texts as databases themselves.[34] In this sense, migrating content into database form is not just about new possibilities but also conforming to the ways that contemporary scholars and other users understand and interpret archival information. Before this work can be done effectively, metadata must be consistent.

The power of relational databases and the popularity of a metadata management system such as ArchivesSpace can allow archives and their users to draw powerful connections across collections. If legacy data is not consistent, as has been shown in this case study, information professionals and their constituents cannot easily establish connections between thematically linked content. We cannot so easily escape our institutional legacies and our legacy data. What we can do as is continue to

identify our catalogs and finding aids as metadata ready to be transformed and reformatted effectively and consistently.

Conclusion

Our experiences at the NAL demonstrate that the content and form of library and archival descriptions are the direct result of organizational decisions and changes. These can range from evolutions in mission, mandates from parent organizations and technological opportunities. The results can often be messy and deter content interoperability. NAL's example does offer solutions for libraries and archives to mitigate these problems: Although creating a subject-based digital collection, they could consult with subject-area experts while both using automated systems, consider authority control, and perform detailed, item-level research. There may not be a universal model for reconciling different or conflicting metadata formats. Instead, we urge institutions to acknowledge their histories and recognize the arbitrary nature of descriptive infrastructures as they build new ones, while pursuing formal and ad-hoc conversion mechanisms to promote interoperability. If they are successful, the results will be powerful. Digital technology enables information professionals to create new digital collections to better serve its constituencies and present its collections in exciting new contexts.

Acknowledgments

We would like to acknowledge Cathy Alessi, Karen Donato, Holly McPeak, Kathryn McMurray, Wayne Olson, and Susan Welsh for sharing their knowledge of the history of the National Agricultural Library and dietary guidance promotion. We also wish to thank Jesse Johnston and Adam Kriesberg of the University of Maryland's College of Information Studies for reading earlier drafts of this article.

Notes

1. Richard Pearce-Moses, *A Glossary of Archival and Records Terminology* (Chicago, IL: Society of American Archivists, 2005).
2. Society of American Archivists, *EAD Application Guidelines for Version 1.0*, accessed January 14, 2015, http://www.loc.gov/ead/ag/agappf.html.
3. Carole L. Palmer, "Thematic Digital Collections," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, and John Unsworth (Oxford, UK: Blackwell, 2004), accessed January 14, 2015, http://digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-4-5&toc.id=0&brand=9781405103213_brand. John Unsworth, "Thematic Research Collections" (paper, Modern Language Association Annual Conference, Washington, DC, December 28, 2000), accessed January 14, 2015, http://people.brandeis.edu/~unsworth/MLA.00/.
4. Kenneth M. Price, "Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?" *Digital Humanities Quarterly* 3, no. 3 (2009), accessed January 14, 2015, http://www.digitalhumanities.org/dhq/vol/3/3/000053/000053.html.

5. Liz Muller. "Publishing the Archive: Definitions and Typologies." *Archive Journal* 4 (2014), accessed March 4, 2015, http://www.archivejournal.net/issue/4/archives-remixed/publishing-the-archive-definitions-and-typologies/.
6. David Bearman and Margaret Hedstrom, "Reinventing Archives for Electronic Records: Alternative Service Delivery Options," in *Electronic Records Management Program Strategies*, ed. Margaret Hedstrom, (Pittsburgh, PA: Archives and Museum Informatics, 1993), 82–98.
7. Anne J. Gilliland-Swetland, *Enduring Paradigm, New Opportunities* (Washington, DC: Council on Library and Information Resources, 2000). Jefferson Bailey, "Disrespect Des Fonds: Rethinking Arrangement and Description in Born-Digital Archives," *Archive Journal* 3 (2013), accessed January 14, 2015, http://www.archivejournal.net/issue/3/archives-remixed/disrespect-des-fonds-rethinking-arrangement-and-description-in-born-digital-archives/.
8. Margaret Hedstrom and John Leslie King, "Epistemic Infrastructure in the Rise of the Knowledge Economy," in *Advancing Knowledge and the Knowledge Economy*, ed. Brian Kahin and Dominique Foray (Cambridge, MA: The MIT Press, 2004), 113–134.
9. Ibid., 124.
10. David Bearman, "Archival Methods," *Archives and Museum Informatics Technical Reports* 3, no. 1 (1989), accessed January 14, 2015, http://www.archimuse.com/publishing/archival_methods/.
11. Daniel Pitti, "Encoded Archival Description: The Development of an Encoding Standard for Archival Finding Aids," *American Archivist* 60, no. 3 (1997): 268–83.
12. Ibid., p. 283.
13. Michelle Light and Tom Hyry, "Colophons and Annotations: New Directions for the Finding Aid," *American Archivist* 65, no. 2 (2002): 216–30.
14. Wendy N. Duff and Vern Harris, "Stories and Names: Archival Description as Narrating Records and Constructing Meanings," *Archival Science* 2, nos. 3–4 (2002): 263–85.
15. Elizabeth Yakel, "Archival Representation," *Archival Science* 3, no. 1 (2003): 1–25.
16. Ibid.
17. For example, see Elizabeth Yakel, "Encoded Archival Description: Are Finding Aids Boundary Spanners or Barriers for Use?" *Journal of Archival Organization*, 2, no. 1/2: 63–77, doi:10.1300/J201v02n01_06.
18. J. Gordon Daines and Cory L. Nimer, "Re-Imagining Archival Display: Creating User-Friendly Finding Aids," *Journal of Archival Organization* 9, no. 1 (2011): 4–31, doi:10.1080/15332748.2011.574019.
19. Jane Zhang, "Archival Representation in the Digital Age," *Journal of Archival Organization* 10, no.1 (2012): 45–68, doi:10.1080/15332748.2012.677671.
20. Naomi Dushay and Diane I. Hillman, "Analyzing Metadata for Effective Use and Re-Use," accessed January 14, 2015, http://ecommons.library.cornell.edu/bitstream/1813/7896/1/501_Paper24-1.pdf.
21. Jung-ran Park and Yuji Tosaka, "Metadata Creation Practices in Digital Repositories And Collections: Schemata, Selection Criteria, And Interoperability," *Information Technology and Libraries* 29, no. 3 (2010): 104–16.
22. Junli Diao and Mirtha A. Hernández, "Transferring Cataloging Legacies into Descriptive Metadata Creation in Digital Projects: Catalogers' Perspective," *Journal of Library Metadata* 14, no. 2 (2014): 130–45, doi:10.1080/19386389.2014.909670.
23. Andrew Wilson, "How Much is Enough: Metadata for Preserving Digital Data," *Journal of Library Metadata* 10, no. 2–3 (2010): 205–217. doi: 10.1080/19386389.2010.506395.
24. Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman, "Science Friction: Data, Metadata, and Collaboration," *Social Studies of Science* 41, no. 5 (2011): 667–90, doi:10.1177/0306312711413314.

25. Mark Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *American Archivist* 68, no. 2 (2005): 208–63. Christine Weideman, "Accessioning as Processing," *American Archivist* 69, no. 2 (2006): 274–83.
26. Carrie Hintz, "The Legacy Landscape," *Chaos ->Order* (blog), accessed January 14, 2015, http://icantiemyownshoes.wordpress.com/2014/02/24/the-legacy-landscape/. Cassie Schmitt, "History and Politics," *Chaos ->Order*, accessed January 14, 2015, http://icantiemyownshoes.wordpress.com/2014/03/11/history-and-politics/.
27. Clive M. McCay, ed., *Notes on the History of Nutrition Research* (Bern, Germany: Hans Huber Publications, 1973).
28. Elaine N. McIntosh, *American Food Habits in Historical Perspective* (Westport, CT: Praeger, 1995).
29. An Act to Establish a Department of Agriculture, Public Law 37-72, *U.S. Statutes at Large* 12 (1863): 387–388.
30. United States Department of Agriculture Library, *Annual Report of the Librarian* (Washington, DC: Government Printing Office, 1899).
31. United States Department of Agriculture Library, *Annual Report of the Librarian* (Washington, DC: Government Printing Office, 1906).
32. United States Department of Agriculture, Center for Nutrition Policy and Promotion. "Guide for Authors and Reviewers" (2012), accessed March 24, 2015, http://www.cnpp.usda.gov/sites/default/files/dietary_guidelines_for_americans/GuideForAuthorsAndReviewers.pdf
33. Eric Forte, Cassandra J. Hartnett, and Andrea L. Sevetson. *Fundamentals of Government Information: Mining, Finding, Evaluating and Using Government Resources* (New York, NY: Neal Schuman, 2011).
34. Lev Manovich, *The Language of New Media* (Cambridge: The MIT Press, 2001). Ed Folsom, "Database as Genre: The Epic Transformation of Archives," *PMLA* 122, no. 5 (2007): 1571–9.