The Blended Cataloguer in the Post-Digital Library
Data Curator, Knowledge Creator, Information Policymaker

Kimmy Szeto

Doctor of Philosophy

City, University of London

Department of Library and Information Science

March 2024

Table of Contents

List of Tables

List of Figures

Declaration

I, Kimmy Szeto, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This dissertation consists of a critical commentary anchored on the portfolio of publications consisting of six peer-reviewed papers selected for their breadth of coverage on library cataloguing issues and also for their evolutionary and futuristic outlook. The overall project traces the transformation of the work of the professional library cataloguer through examining advances in the past decade as documented in the prior publications, and charts the expansionary trajectory of the profession in light of the ongoing digital transition in libraries. The critical commentary places cataloguing, the catalogue, the cataloguer, and catalogue librarianship in the evolutionary framework for library services, described in Michael Buckland's 1992 publication *Redesigning Library Services: A Manifesto*, and subsequent extensions of this framework by other authors. Into this framework, a new "Blended" state, omitted as transitory between "Paper" and "Electronic" in Buckland's framework, is inserted, and a new professional identity of the "Blended Cataloguer" is developed to take on the roles of data curator, knowledge creator, and information policymaker. A unified practice surrounding these new roles, the "Decomposition-Assembly Approach" is developed to describe a data processing methodology incorporating both traditional and digital library cataloguing skills. Beyond complementing the new Blended Cataloguer professional identity, this approach is also shown to represent a practical approach to "enrich and filter," a theoretical model for digital library metadata developed by Alemu and Stevens (2015). The Blended Cataloguer practising the Decomposition-Assembly Approach represents a radical shift from the traditional standard-based practise, and provides a new evolutionary framework and practical model for current library cataloguers and metadata workers that will enable the development of foundational information infrastructure for future library services and provide leadership in shaping the broader information ecosystem.

Acknowledgements

Writing a dissertation is no small feat. It takes a village from start to finish. I am incredibly fortunate to be part of the community at City's Library and Information Science Department and the supportive cohort of doctoral researchers. In particular, I would like to express my heart-felt thanks to Prof. David Bawden for guiding and focusing my writing with amazing patience and a healthy dose of humour, to Dr. Lyn Robinson for addressing the holistic wellbeing in the process, and to both of them for believing in the significance of this project. I also thank Dr. Mark McKnight and Dr. Arthur Downing who, at the start, vouched for me as a researcher, Veronica Alzalde Wells and Nara Newcomer who, after enduring the reading of a bumpy draft, asked probing questions which greatly refine the text, and to Dr. Minna Vuohelainen and Dr. Joseph Thomas who helped shepherd the project to the finish line.

Further back in time, I would like to thank John Hyslop and Judith Todman at the Queens Public Library for their challenging assignments which planted the seed of this area of research, Constantia Constantinou at Maritime College, State University of New York who afforded me time and space to pursue this line of intellectual query, Dr. Arthur Downing at Baruch College, City University of New York who nurtured my professional growth, the many members of the Music Library Association in the United States who provided me a platform to speak and finetune ideas, Deborah Compana, former editor of the journal *Notes*, who unwittingly showed me an effective editing process which greatly strengthened the writing of this dissertation, and Terra Merkey, Veronica Alzalde Wells, and Laikin Dantchenko whose passion for the profession and their care for those in the profession strengthened my resolve to embrace an inclusive and expansionary library future.

Finally, I thank my wife Dr. Anne Lovering Rounds for her unwavering support throughout this journey (especially the late night proofreading marathons) and my daughter Katherine Calliope for providing the daily rhythm and laughter so necessary for staying healthy and positive through this process.

Chapter 1. Introduction

Introduction to the Topic and its Significance

This submission for the Ph.D. by Prior Publication traces the transformation of the work of the professional library cataloguer as the library community adopts linked data technologies and shares library data in the open, global information space.

Cataloguing plays a central and critical role in library operations. Traditionally, as a technical operation, the primary task of the library cataloguer is to produce a library catalogue, which describes the holdings in the library's collections. The catalogue would be consulted by the technical staff for other operations such as material acquisition, processing, and shelving, as well as by the library services staff to meet user questions and requests. The cataloguer treads the space between library staff and library users through the creation of the library catalogue, often having to make judgement about adhering to cataloguing rules and standards while addressing internal and external service demands. The design, content, and presentation of the library catalogue can have a significant impact on technical operations and library services. The cataloguing profession is, therefore, inextricably tied to materials and services.

The digital revolution ushered in a period of accelerating change for libraries in the past half century. The advent of the computer automated library technical operations while the internet enabled an open, global information space. In addition to traditional, paper-based materials and in-person services, libraries now collect electronic materials and provide services online. The expanded scope calls for expanded roles for the library catalogue. With changes in the library catalogue come changes in library cataloguing, which has been finding its way to a new equilibrium between materials and services.

Researching the transformation of the library cataloguer addresses the lack of work on conceptualising the cataloguing profession in this period of rapid change. The openness of the global information space introduced uncertainties over the scope of work and the space the cataloguing profession occupies in libraries. Even how to refer to the profession itself has been called into question. The aim of this dissertation is to place library cataloguing in the context of evolving library collections and services through a close examination of the transformation of the scope of cataloguing work “on the ground.” The prior publications reveal that this transformation not only requires applying cataloguing theories and skills to a new technological landscape, but also requires claiming leadership in standardisation and outreach so that librarians will be able to expand their authoritative stature in the information domain. The traditional library cataloguer will see an expansionary transformation of scope into areas of expert data curator, knowledge creator, and information policymaker.

The most significant contribution to knowledge of the study in this dissertation is in redefining the scope of library cataloguing, appreciating the impact of the library catalogue, affirming the value of library cataloguers, and charting a new course for the cataloguing profession in the twenty-first-century post-digital revolution library. This research is most relevant to cataloguers, library institutions, as well as workers in adjacent fields who collaborate with librarians. A clear understanding of the profession can guide individual cataloguers, aspiring and seasoned, on their paths toward professional development, as well as help institutions create sensible plans for the organisation of library operations, hold realistic expectations of their cataloguing staff, and develop effective and efficient collaborations between units within and across institutions.

Introduction to the Methodology

This critical commentary addresses two research questions through an analysis of the transformation of the library cataloguing profession, as exemplified in the prior publications, in the framework of evolution of library services described and anticipated in *Redesigning Library Services: A Manifesto* by Michael Buckland (1992) toward dematerialization. The narrative will follow three themes: cataloguing of library materials, the library catalogue as a service tool, and the library cataloguer as a driver of library innovation and leadership. While the first two themes are directly based on Buckland’s work, the third theme is the product of a broadening of Buckland’s framework based on the 2014 essays by Brett Bonfield, Hugh Rundle, and Megan Hodge in the edited volume *Planning Our Future Libraries: Blueprints for 2025*. At the intersection of these two publications and the prior publications is the trajectory of library’s transformation in an era of rapid technological change. The prior publications elucidate a latent dialogue between these two critical perspectives, published twenty years apart. The connection between the two will become the focal point of the later chapters, guiding this critical commentary to its conclusion.

The body of prior publications reveals an expansionary trajectory of the library cataloguing profession which has evolved alongside library services. Discussions in this critical commentary are anchored on these prior publications. Since these prior publications were published in specialised journals in the field, historical background and supplementary information are referenced and inserted into the narrative as appropriate to complete the picture for the more general reader.

While the direction of library collections, services, and the catalogue have been a frequent topic in research literature, the cataloguing profession has rarely been discussed or theorised in this context. By addressing the two research questions through this analysis, it is hoped that this dissertation as a whole will contribute crucial directions to the library cataloguing profession in the twenty-first century.

Structure of the Dissertation

This submission for the Ph.D. by Prior Publication consists of six peer-reviewed publications and this critical commentary.

- Critical commentary
- Reference list
- Prior publications [Papers 1-6]

The prior publications

In chronological order of publication, the papers submitted are as follows:

[Paper 1] Szeto, K. (2011) ‘Digitizing everything? [Part I] The launch of the digital imaging program at Queens Library’, and ‘Part II: Piloting metadata creation’. *Journal of the Library Administration & Management Section*, 7(2), pp. 5-15, and 8(1), pp. 31-49.

[Paper 2] Szeto, K. (2013) ‘Positioning library data for the Semantic Web: recent developments in resource description’. *Journal of Web Librarianship*, 7(3), pp. 305–321. DOI: 10.1080/19322909.2013.802584

[Paper 3] Szeto, K. (2017) ‘The mystery of the Schubert song: the linked data promise. *Notes, the Quarterly Journal of the Music Library Association*, 74(1), pp. 9-23. DOI: 10.1353/not.2017.0071

[Paper 4] Szeto, K. (2018) ‘Metadata standards in digital audio’, in Khosrow-Pour, M. (ed.), *Encyclopedia of Information Science and Technology*. 4th ed. Hershey, Pennsylvania, USA: IGI Global, vol. 8, pp. 6447-6463. DOI: 10.4018/978-1-5225-2255-3.ch560

[Paper 5] Szeto, K. (2019) ‘The roles of academic libraries in shaping music publishing in the digital age’. *Library Trends*, 67(2), pp. 303-318. DOI: 10.1353/lib.2018.0038

[Paper 6] Szeto, K. (2022) ‘Ontology for Voice, Instruments, and Ensembles (OnVIE): revisiting the medium of performance concept for enhanced discoverability’. *Code4Lib Journal*, 54. Available at: https://journal.code4lib.org/articles/16608

Statement of authorship

All the above papers submitted as prior publications were solely authored by the same author of this critical commentary and this overall dissertation.

References to the prior publications

Each paper is referenced throughout this critical commentary by its number ([Paper 1], [Paper 2], etc.) References to these papers follow the pagination of this overall dissertation, not those in the papers as published.

Publication history

[Papers 1, 2, 3, 5 and 6] appeared in peer-reviewed journals. [Paper 4] was the revised and updated edition which was peer reviewed a second time to become published as an entry in an encyclopaedia. This paper was subsequently re-published as a chapter in the edited book *Advanced Methodologies and*

Technologies in Media and Communications by IGI Global in 2019 (DOI: 10.4018/978-1-5225-7601-3.ch020).

Provenance of the critical commentary and the prior publications

The six papers of the prior publications were collected to represent the research outputs of my career as a member of the academic faculty in a library setting.

[Paper 1], in two parts, represents a retrospective of my work on developing a digitisation programme as a practitioner at the Queensborough Public Library (now renamed Queens Library), one of the three public library systems serving the Borough of Queens in New York City. The first part establishes the philosophical foundation and qualifications of the cataloguer for such a programme. The second part is a case study of a proof-of-concept demonstration of a novel, non-traditional library service by presenting the catalogue of digitised materials online in cartographic form. The work included establishing the workflow for cataloguing a set of archival photographic materials and their digitised images, which involved developing novel data conversion, automation techniques, and staff training in the cataloguing department and in the Long Island Division, the archives unit where the materials and digitisation equipment were housed.

[Paper 2] grew out of my years of sustained effort to introduce the semantic web and linked data concepts at plenary and paper sessions at the Music Library Association conference in the United States. Feedback from these sessions revealed a great demand from the cataloguing community to learn more about the “what” of cataloguing activities and “where” the library catalogue is situated in the broader information space. This paper builds a framework in which cataloguers gradually become data curators in a more generalised information landscape.

[Paper 3] began as my response to a bibliographic reference situation and the realisation that information seekers often grapple with the disconnect between data existing everywhere but nowhere to be found. This paper examines the “how”—given the library catalogue, how could the cataloguing profession propel the library’s move to connect bibliographic data to the open web? Through the example of medium of performance, which would later be revisited in [Paper 6], this paper theorises data modelling as a solution, a radical move in a profession where cataloguing rules and standards have been relatively stable.

[Paper 4] can be considered a technical supplement to Paper 2, offering a comprehensive survey of the state of the art of embedded metadata in digital audio. As digital materials grow in library collections, cataloguers can increasingly benefit from harnessing the richness of metadata in the many formats, as well as the awareness of metadata created at various steps—the “when”—of the production chain and in the hand of the consumer. The ability to manipulate these embedded metadata affords the processing of these metadata as the digital audio contents interact with bibliographic and discovery systems, and afford the potential to create knowledge through automated, self-generating catalogues.

[Paper 5] theorises the policy role of the academic libraries in the post-digital revolution information ecosystem and demonstrates the convergence of production, publication, resource description, discovery, access, and user manipulation through proposing a set of standard functional requirements for embedded metadata in electronic music scores. The paper advocates for the involvement of libraries in the standardisation of the entire life cycle of electronic music scores, a role well-suited for cataloguers.

[Paper 6] provides the technical realisation of the data modelling solution to library data pertaining to medium of performance. Through a thorough redesign of the MARC21 382 field to an extensible linked data structure, the paper examines fundamental assumptions and addresses longstanding conceptualisation in traditional library cataloguing, and advocates for cataloguers to transcend these constraints in future work.

Influences of Prior Publications and Future Research

These prior publications have become foundational for a number of research directions. [Paper 2], [Paper 3], and [Paper 5] were discussed in the “literature review” of over a dozen articles and book chapters; [Paper 4] was cited in a patent on time marker metadata in digital audio; [Paper 2], [Paper 3], [Paper 5] and [Paper 6] appeared in systematic literature reviews of significant research in publications in English, Portuguese, Catalan, and Chinese (for example: Gaitanou, et al. (2022); Patrício, Cordeiro and Ramos (2021), Wang and Yang (2018), and Ke, Yuan and Yang (2019)).

The crosswalk in [Paper 1] ultimately was expanded to Appendix A in the co-authored paper *Music discovery requirements: A guide to optimizing interfaces*

(Newcomer *et al*., 2013), an influential publication on shaping OPAC search and display of music resources in the past decade. The workflow in [Paper 1] endures to this day in the Queens Library and the metadata created in the paper found its way into a new discovery layer that is still in use today (Queens Public Library, 2019). Unfortunately, the KML cartographic implementation has been rendered moot by Google.

While the majority of the citations appeared in publications related to linked data, semantic web ontologies, and music linked data, the significance of a vertically interoperable music metadata proposed in [Paper 5] was discussed in two dissertations, one for a business application related to royalties collections (Hawkins, 2021), another on the topic of music information retrieval (Peetz-Ullman, 2020) concerning the potential application for a music read-along display on electronic music scores through the OPAC. This application represents a subset of capabilities developed by engineering researchers in the field of temporal hypermedia with vast applications in the nascent field of digital musicology (such as the Music Encoding and Linked Data group, see Page, Lewis and Weigl, 2019). The author of the dissertation continued to mention new demands for searching music by any imaginable facet of music, from watermarks on the paper to the composer's gender. The library data modelling framework proposed in [Paper 6], published two years later, would actually make this possible. In fact, research in [Paper 6] contributed to an internationally coordinated effort to enhance discoverability of music materials on the web (LinkedMusic, 2023).

Research Questions

In this critical commentary, I aim to address these overarching research questions:

1. In what ways has the cataloguer's role transformed and in what ways will it continue to transform as the library community adopts linked data technologies and shares library data in the open information space?
2. In what ways and to what extent will this transformation necessitate revisions and innovation of the cataloguer's knowledge, skills, and tools?
3. In light of these transformations, can a practical model of cataloguing be developed?

Structure of this Critical Commentary

Chapter 1

Chapter 1 (this chapter) introduces the research questions, the bibliography of the prior publications, and the structure of this critical commentary and the dissertation as a whole.

Chapter 2

Chapter 2 introduces the theme and thread of this critical commentary. The theme focuses on the persistent period of coexistence between paper and electronic library materials and services. The term “Blended Library” is introduced to describe this state, and, correspondingly, the “Blended Cataloguer” who operates in this state. The thread that continues through the evolution of libraries is cataloguing, as it continues to be a core operation in libraries and a significant area of librarianship. Along this thread, the “Decomposition-Assembly Approach” of cataloguing is introduced. I argue that this generalised metadata approach is well suited for the Blended Library, and I advocate embracing the approach as a core part of the professional identity of the Blended Cataloguer.

Chapter 3

Chapter 3 delves into the interactions between cataloguing practices and the nature of library materials. The term Blended Library is placed into the historical framework of library services. Issues of cataloguing blended paper and electronic materials are explored in the context of the Decomposition-Assembly Approach.

Chapter 4

Chapter 4 places the Decomposition-Assembly Approach in the context of the functions and services provided by the library catalogue. After a brief history of the library catalogue, the chapter’s focus turns to situating the library catalogue through technological developments that necessitated a radical transformation of approach to cataloguing. The chapter comes full circle with a discussion on the novel forms and functions of the library catalogue produced by the Decomposition-Assembly Approach.

Chapter 5

Chapter 5 places the Blended Cataloguer in the context of the cataloguing profession through an examination of education, training, and position descriptions of library cataloguers. The characteristics of the Blended Cataloguer are delineated in contrast to the “Blended Librarian” which is a term that describes the public service counterpart.

Chapter 6

Chapter 6 presents a synthesis of the role of the Blended Cataloguer within the Blended Library environment and discusses the future professional landscape of the library cataloguing profession. Several emergent areas are described, and their implications on the future direction of library services are discussed.

Chapter 7

Chapter 7 returns to the research questions posed in Chapter 1.

Reference list

The reference list includes works cited in this critical commentary. Bibliographies referenced in the prior publications appear at the end of each prior publication.

Chapter 2. Themes and Threads

Framework

In his 1992 publication *Redesigning Library Services: A Manifesto*, Michael Buckland summarised the state of automation of library technical operations and anticipated the stages through which they would go from being paper-based to electronic. He described the underlying concerns of library technical operations—terminology, scale of operation, and technical refinements—from the previous 100 years as “strikingly contemporary.” In the subsequent three decades, libraries have, indeed, moved toward dematerialisation, and these concerns are as relevant today.

Technological advances thrusted library cataloguing into a blended space between library materials and library services, between library workers and users, in which the paper and the electronic coexist. Occupying this blended space, the cataloguing profession is a bellwether of libraries and librarianship, both as an indicator and predictor of change.

The prior publications presented in this dissertation exemplify the profession of library cataloguing in this blended space through detailing the scope and workflow of the library cataloguer, from library materials to the library catalogue. Collectively, the six prior publications encapsulate the transformation of the library cataloguing profession following a long, stable period until about a decade ago when digital items began to be systematically collected by libraries, and when library services began to be systematically offered on digital platforms.

As such, this critical commentary will be organised by themes, the first two focusing on library materials (Chapter 3) and library services through the catalogue (Chapter 4), and the third covering the librarian (Chapter 5), and librarianship (Chapter 6). The two common threads of blendedness of cataloguing and the Decomposition-Assembly Approach to library data will appear throughout.

Threads

Thread 1: The blended space

Automation in libraries has existed since the 1960s, although rapid advance in the application of computer technologies in library operations did not begin in earnest until the 1990s. Buckland, writing at this time, anticipated that libraries would substantially move from the “Paper Library” where materials and technical operations were both performed on paper or paper-like mediums, to the “Automated Library” where technical operations were largely assisted by computer technology while materials remained “overwhelmingly” on paper mediums. In this framework, he further envisioned the future “Electronic Library,” where both library materials and services would take on dematerialised forms. Nonetheless, he predicted that, more realistically, the Automated Library and Electronic Library would coexist through a transitional period in a blended form.

Today, libraries do, in fact, exist in this blended form. Libraries have collected large amounts of electronic materials while showing few signs of wholesale abandonment of paper collections. Libraries have expanded their electronic services while keeping their physical space, albeit repurposing it. The prior publications in this dissertation reveal that the cataloguing profession has adapted to this persistent blended state by incorporating practices that address library materials and services in electronic and networked environments while maintaining ties to the paper materials and catalogue.

By connecting these adaptations, this critical commentary draws out the recent transformation and future prospect of the library cataloguing profession and argues for the insertion of the “Blended Library” to Buckland’s framework. With this conceptualisation, connections will be drawn between the Blended Library and the “Participatory Library” proposed by Bonfield in 2014, as well as between the “Blended Cataloguer” and the “Blended Librarian,” a term coined for academic librarians by Steven Bell and John Shank in 2004.

	Library Materials	Technical Operations	Cataloguing Mode	Library Catalogue
Paper Library	Paper	Paper	Traditional	On paper
Automated Library	Paper	Computer	Traditional	On paper, also accessed online
Blended Library	Some paper + overwhelmingly electronic media	Computer	Blended	Based on paper, accessed online, and produced online
Electronic Library	Electronic media	Computer	Data Curation	Produced online

Table 1. Technological Bases of Library Operations and Materials and the Cataloguing and Catalogue that Support Them. This table is expanded from Table 1 in Michael Buckland's *Redesigning Library Services: A Manifesto* (1992). Additional shaded cells are added to the original table which is unchanged.

Thread 2: The Decomposition-Assembly Approach to cataloguing

I brought decomposition and assembly techniques to library data from my previous experience before entering the library profession, as a scientific programmer. Widely used in computer programming, computer aided design, and manufacturing, these techniques emphasise modularisation of data structures and functional requirements. In library applications, examples of decomposition can be found in the modules of integrated library systems (acquisition, cataloguing, circulation, finance, etc.), whereas assembly describes processes that gather data from these modules to create reports. The challenge in applying these techniques to cataloguing is that the standardised structure of cataloguing practice and the catalogue itself cannot be changed as quickly as the variable demands of library services, which results in a anticipation-and-catch-up dynamic to satisfy functional requirements that are susceptible to more frequent changes.

I work in the library as a cataloguer, rather than a software engineer, systems engineer, or a project manager. However, the body of prior publications reveals the benefits, and necessity, of employing decomposition techniques in processes of cataloguing and metadata work, and assembly techniques to produce the desired output, in the form of a library catalogue or in other physical or digital forms. Some forms of these techniques have already been employed in traditional cataloguing practice, most notably constituting a library catalogue with cards rather than printing its entirety in book volumes. In recent years, this approach has quietly become ever more pervasive in cataloguing and metadata work as the information ecosystem shifts from a document-based landscape to one of semantic linked data, from one for exclusively human consumption to one also catered to machine processing. In this critical commentary, the Decomposition-Assembly Approach employed in the prior publications is highlighted as a necessity in support of the expansionary trajectory of the Blended Cataloguer.

Themes

Theme 1: Cataloguing blended library materials

"The central purpose of libraries is to provide a service: access to information": so begins Michael Buckland's 1992 *Manifesto*. To provide this service, libraries systematically and purposefully acquired materials that contain information into their collections. Until the 1970s, the primary medium carrying such information was paper: books, periodicals, maps. Material formats such as microforms and audiovisuals formed a small portion. These non-book formats presented in similar physical forms and were offered to library users in a similar manner to their book counterparts. The rise of computer and network technologies has information increasingly stored and transmitted electronically. Libraries followed by providing materials in electronic mediums. While some electronic mediums such as the compact disc possess characteristics of physical objects, many exist purely as digital objects. Some digital objects have physical counterparts, some others are born digital with no physical counterparts.

Digital objects derived from paper materials

In the case where the physical counterparts exist, it would be logical to bring the physical and digital together and catalogue both formats as a unified collection. This represents a blended scenario at the local collection level. [Paper 1] is a case study of early adaptations of the library cataloguer to such a blended collection. The methodology adopted in [Paper 1] leveraged traditional cataloguing practices to extended descriptions and encoding for digital images. The methodology is a novel process that uses the Decomposition-Assembly Approach to produce RDF-style linked data, while accommodating the traditional cataloguing practice of inputting data directly into a MARC record using the then-prevailing AACR2 description standard. The process begins with a custom-built MARC-like cataloguing work form. Descriptive cataloguing data from the work form was then taken apart, merged with additional administrative, geographical and technical metadata, and assembled into Dublin Core, VRA Core, and KML metadata records. This approach gave traditional cataloguers a means to make their first foray into working with digital materials. In addition, this case study was the first in the literature to present a crosswalk between AACR2/MARC, VRA Core, and Dublin Core schemas.

This blended approach presented in [Paper 1] applied the assistance of computers of Buckland's Automated Library to catalogue materials in an Electronic Library. The resulting output enabled the creation of a meaningful set of metadata for a digital collection that are connected to their analogue collection counterpart. The metadata further enabled presentation of the catalogue in the cartographic medium, an innovation at the time that would only become popularised by Google Maps five years later.

Natively digital objects

In the case where the materials are born digital, the cataloguer would benefit from gathering information from the digital materials themselves, just as a traditional cataloguer would derive catalogue descriptions from examining the physical item. While [Paper 1] offered a workflow for cataloguing digital materials, the descriptions were primarily based on their physical counterparts. [Paper 4] scopes out new territories as it looks into metadata embedded in digital materials.

Library cataloguing has traditionally been bibliographic, considering audiovisual materials as physical publications, based on what the cataloguer can see or hear in the contents and can read on the carrier and container. With electronic materials, it was necessary to expand the scope of information gathering in the non-visible parts of the digital carriers. As with bibliographic standards, data formats for audiovisual materials are maintained by standards organisations, which are dominated by equipment manufacturers, with a handful of *de facto* standards made popular by consumer software. These standards provide uniformity for descriptive and technical metadata embedded in the digital files which can then be passed down the production chain and, ultimately, to end users. Since these metadata are already electronic, they are ready to be plugged into automated processes. [Paper 4] raised the awareness of these useful embedded metadata and advocated their use through the Decomposition-Assembly Approach, not only for library cataloguing of blended collections exemplified in [Paper 1], but also for innovative, creative uses that further cultural production.

Theme 2: The Catalogue in the Blended Library

When Buckland published his framework in 1992, the Automated Library was increasingly benefiting from computers in library technical operations, while dealing with paper-based library materials that remained largely unchanged. Griscom (2022) provides a detailed first hand account of library cataloguing operations up to this point. In his account, as cataloguers produced each record, the information was also entered into a central computer database which facilitated the printing of catalogue cards. While this central database was always reached via a staff-only computer network, software was being developed to take advantage of the networked environment for library users to retrieve catalogue card surrogates at designated remote terminals. The late 1980s and 1990s saw the gradual adoption of such remote access by libraries. This new library service was part of the trend toward “self-service, open stacks and public catalogues,” a prediction of twentieth-century library service in Buckland (1992), and this public catalogue would be electronic and eventually termed the online public access catalogue (OPAC).

At this time, the library catalogue, delivered electronically, presented a second layer of surrogates: the electronic surrogates of library catalogue cards, which are themselves surrogates of library materials. Library cataloguing had been designed for use within the library walls, and, with cards, designed to fit within the edges of catalogue cards. Surrounding the site of use and physical limitations created a specialised semantic in the language of library cataloguing. However, the computer screen not only had different dimensions, but also offered a flexibility to display beyond one screen’s worth of information. And, because more information could be displayed, there could be more potential uses for such remote access, more flexible ways of sorting, searching, and filtering the cards. In fact, coded fields had been created in the MARC format to facilitate these functionalities, and cataloguers had already been inputting data in these coded fields in anticipation of the OPAC.

Considering new uses of the library catalogue and the new electronic space where the library catalogue is situated, [Paper 2] offered a re-examination of the design of the library catalogue in this blended environment, where the catalogue was subject to more robust, granular search methods and results are delivered and displayed in a wide variety of formats. This process of displaying search results based on querying individual fields of the catalogue is akin to using the Decomposition-Assembly Approach to create a custom catalogue on each search. To make this possible, the library catalogue as a whole needs to be created in such a way that makes it available to be decomposed. The new conceptual model (FRBR) and resource description standard (RDA), which would become the foundations of the library cataloguing profession, was a significant step towards this capability. [Paper 2] traces the technological convergence that brought cataloguing to this point, and elucidates how the traditional library catalogue could be interpreted in the new FRBR model and the RDA standard, and how MARC fields could be mapped to RDF-style linked data, the fundamental building blocks of the semantic web, thus re-orienting library cataloguing toward a model of "on the web, for the web."

In the course of re-conceptualising the library catalogue, not all areas of descriptions are able to make the leap without significant modification due primarily to conventions based on human practices. Such areas would require individual re-conceptualising and re-modelling from scratch. [Paper 3] examined the various pitfalls of forcing the medium of performance area of description to fit into the existing data model and illustrated why it could not be effectively used in an electronic library catalogue. Through introducing a new conceptualisation for this area of description, [Paper 3] lays down the conceptual design for the library catalogue that adequately provides functionalities that respond to library services in the Blended Library environment, offering a blueprint for building an area of the library catalogue under this new orientation.

Theme 3: The Blended Cataloguer

The scope of the library cataloguer evolves together with library services and available technology. Cataloguing, traditionally a part of technical services departments in libraries, is naturally affected by the technical advances and by the library collections and services. In response to blended library materials and the blended catalogue, the Blended Cataloguer is one who rises to the challenges of the Blended Library, one who operates within the continuous spectrum from the Paper to the Automated to the Electronic Library, and one who possesses the expertise to manipulate library data along this spectrum to create meaningful library services. The Blended Cataloguer is fluent not only with traditional library cataloguing, but is also able to adapt traditional library cataloguing methodologies to lead the design and creation of electronic library catalogues that can be used for generalised library services in an online environment.

The term Blended Cataloguer is coined in this critical commentary as the counterpart of the "Blended Librarian," a term coined by Steven Bell and John Shank (2004) to describe future public service academic librarians. Bell and Shank define the Blended Librarian as "an academic librarian who combines the traditional skill set of librarianship with the information technologist's hardware/software skills, and the instructional or educational designer's ability to apply technology appropriately in the teaching-learning process." The work highlighted in the prior publications reveals that the Blended Cataloguer combines the skill set of the traditional cataloguer with the information specialist's data manipulation skills, and an interface designer's ability to serve data in the appropriate formats to library users and other automated systems. The broad scope of the Blended Cataloguer has the potential to influence, alter, and design the technologies.

[Paper 5] and [Paper 6] are demonstrations of the Blended Cataloguer influencing the technological landscape with direct impacts on library services. [Paper 6] implements the conceptual design in [Paper 3] and develops a new data model for the medium of performance area of description. [Paper 6] reveals medium of performance as an area of divergence between bibliographic description and user practices. The Blended Cataloguer would be in a position to recognise such a divergence and initiate the reflective process of introducing modifications to library cataloguing such that library data may continue to be a useful service tool.

Cycles of reflection and redesign could lead to developments of new technological standards, a role librarians have generally shied away from. The research in [Paper 5] participated in the early development of the electronic music score when it was an emergent material format. The expertise of a Blended Cataloguer is shown to be valuable for creating new definitions, functional requirements, and the flow of data through the life cycle of these digital objects. [Paper 5] also shows that in so doing, the Blended Cataloguer enables expansion of library services to include technology services that, with electronic materials, necessarily accompany the information service.

The scope of the Blended Cataloguer is no longer limited to compiling descriptions of physical library items into a catalogue. Library cataloguing today takes the lead in defining how electronic materials are formed, disseminated, and integrated into library users' knowledge environment. This transformation of scope harkens back to the vision of the trajectory of the cataloguing profession put forth in [Paper 2]: from recording and transcribing to creating and curating, with the cataloguer's expanded role as policymaker in areas of information management and user experience. With the ability to manage the life cycle of data, [Paper 5] anticipates the future library as more than a service point and becoming a platform. This vision takes the general purpose of the library beyond "access to information" as stated in Buckland's *Manifesto*, and to a space fostering human creativity and innovation.

Chapter 3. Cataloguing the Blended Library

Introduction

In his 1992 *Manifesto*, Michael Buckland predicted that library services would evolve through stages, from the “Paper Library” to the “Automated Library,” and, finally, to the “Electronic Library.” The Paper Library represents a stable period from antiquity to about the 1970s. In this period, libraries systematically and purposefully acquired book collections. Library services evolved to its modern form in the second half of the 19th century, primarily involving transacting information and materials in library collections. In Buckland’s view, collecting and providing services comprise the defining aspects of libraries, and this bipartite model is reflected in scholarly literature.

Yet cataloguers work between these spaces. Cataloguing produces the library catalogue that serves to organise library collections. This very organisation of the collections enables services to be performed. Technical innovations in the cataloguing and physical implementation of the catalogue have made possible standardisation and cooperation between libraries, which, in turn, made possible the trend towards self-service, open stacks, and library catalogues for public use.

From this starting point, Buckland theorised the evolution of library services in times of technological change. The Automated Library incorporates computer technology to augment technical operations while library materials and services remained “overwhelmingly on paper and paper-like media,” whereas in the future Electronic Library, both library materials and services will inhabit a dematerialised form. This trajectory notwithstanding, he concedes that paper materials are unlikely to disappear entirely for quite some time.

The term “Blended Library” is used in this critical commentary to reflect this liminal period in which library collections and services exist in a blended state. This chapter focuses on the challenges in cataloguer’s work during a period when library materials became rapidly blended.

(The term “Blended Library” first appeared in literature in Heilig, Rädle, and Reiterer (2011) to refer to the application of conceptual blending, a theory of cognition, in the design of multimedia, multi-device, tangible user interfaces for library spaces. This usage of the term here is sufficiently distinct.)

The Bibliographic Approach to Cataloguing

Traditional cataloguing practices in the Blended Library

Traditional cataloguing, as performed in the period before computerisation, is a technical operation where library holdings are recorded, described, and indexed (see definition of "Catalog" in Young and Belanger, 1983). The ultimate output of this operation is the library catalogue, which consists of entries that function as surrogates of individual items in the library. These surrogates help libraries maintain organisation of library materials, and allows library users to learn about the library's holdings and to retrieve items.

Cataloguing operations can be roughly divided into two parts: in the broader sense, the maintenance of the library catalogue, which also includes classification, subject analysis, access points, and authority work; in the narrower sense, the creation of bibliographic description for individual entries in the catalogue. Further discussion on cataloguing in the broader context of the library catalogue appears in the next chapter. This chapter traces the cataloguer's role in the creation of bibliographic records in the traditional cataloguing setting and adaptations to the introduction of electronic mediums.

Historically, library collections were built primarily on paper materials, and library services were also transacted on paper. For the purpose of creating surrogates for the library catalogue, the concept of the bibliographic item was developed, focusing on certain characteristics of the item for the catalogue entry. Traditionally, the bibliographic item was synonymous with the physical book. This centuries-old mindset is reflected in the standards and practices of traditional cataloguing.

Due to this stability, as cataloguing developed and evolved, practices were codified and standardised around the cataloguing of the book medium. When computers began to enter library operations in the 1960s, automation reduced repetitive data entry and facilitated catalogue printing, while the bibliocentric approach of cataloguing continued. Library collections continued to consist of paper-based physical materials, and cataloguing continued to focus on bibliographic control, where library materials are described and tracked at the individual item level.

When electronic materials began to be introduced, cataloguing practices gradually fit them into the book cataloguing framework, extracting book-like characteristics from electronic resources and digital materials, as had been the earlier practice for materials in other mediums such as microform and cartographic materials. The emphasis on maintaining bibliographic control by standards maintenance agencies such as the Library of Congress persisted into the mid-2000s amidst the proliferation of electronic resources (Library of Congress, 2005), primarily aided by increased cooperative cataloguing (Ruschoff, 2020).

Maintaining the bibliographic approach to cataloguing reflects the scale of the inertia of the Paper Library. The conceptual model for describing and organising library resources at the time, Functional Requirements for Bibliographic Records (FRBR), published in 1998, represented the recognition of the increased prominence of electronic and multimedia material formats and an attempt to provide a more flexible framework that separated bibliographic items into four layers of abstraction: work, expression, manifestation, and item. While this model redirected the emphasis of cataloguing from the “item in hand” to the four FRBR entity layers and the relationships between them, the end goal continued to be “bibliographic records.” Cataloguers would continue to create traditional library catalogues, and the FRBR model would continue to direct cataloguers to extract book-like characteristics for creating catalogue entries, albeit from a wider range of material formats via more flexible paths.

This retrofitting seemed to be a reasonable step for electronic materials such as ebooks, electronic journals, and digitised materials, which have paper counterparts. For example, Copeland (2002) examined specific challenges of cataloguing digital representations of books using existing cataloguing rules. A similar type of study was presented in [Paper 1] where monograph cataloguing practices were retrofitted for cataloguing photographic prints, photographic plates, and their digitised images. This type of retrofit cataloguing practice of hybrid paper-electronic bibliographic items spurred the development of methodologies and practices that form the basis from which the Blended Cataloguer is developed in later chapters, and become defining features of the cataloguing profession in the Blended Library.

Standardisation of bibliographic description

Bibliographic description refers to the elements of information from a bibliographic item that are transcribed, recorded, and created for the bibliographic record. Traditionally, this work has been carried out by cataloguers trained specifically for the task.

In the course of creating bibliographic descriptions, the cataloguer, ideally, ensures descriptions are comprehensible and consistent across a large number of bibliographic items, across collections, and, in modern practice, across libraries. The work falls into two main areas: bibliographic control, where the cataloguer follows prescriptive rules for transcribing and recording these descriptions so that they consistently refer to the same item, and authority control, where the cataloguer ensures terms and phrases are consistently used across multiple items.

Neither bibliographic control or authority control is a trivial task. But, even though cataloguing has been practised for centuries, practices varied from cataloguer to cataloguer prior to the mid-19th century. Professionalisation of the library workers and changes in communication and information technology over time have increased cooperation between cataloguers and between libraries. Cooperative efforts to maintain this intellectual work of cataloguing requires standardised practices. Traditional cataloguing, therefore, places a strong emphasis on standards development and rule interpretation.

Descriptive cataloguing standards

The library cataloguer not only has to be concerned with describing bibliographic items, but also with the parts of the catalogue in which the descriptions should appear. Cataloguing standards come in two main types: content standard and encoding standard. The content standard governs what information is extracted from the resource, the manner in which information is extracted, and the form of text that ultimately appears in the catalogue; the encoding standard prescribes the structure in which the extracted information is arranged. In bibliographic description, standardisation concerns the content in the description—how to arrive at the description. For catalogue entries, the standardisation of concern is the format—where the description goes within the space of each entry. In the 1990s, when library cataloguing had already been computerised for three decades, the predominant standards were AACR2 (Anglo-American Cataloguing Rules, second edition) for content and MARC (machine-readable cataloguing) for encoding. A new content standard, RDA (Resource Description and Access), was published in 2010 and officially replaced AACR2 in 2013. Overviews of specific areas of cataloguing such as descriptive cataloguing can be found in Dobreski (2021), authority control in Wiederhold and Reeve (2021). A holistic history of standardisation of cataloguing practices has been detailed in Miksa (2021).

[Paper 2] sets this holistic history against the backdrop of the emergence of the semantic web and discusses the ramification of applying the traditional cataloguing skill set to nascent linked data technologies that were anticipated to become the informational backbone of the future. The introduction of electronic library materials and the emergence of an electronic information landscape strained the longstanding bibliographic standards. [Paper 2] offers an analysis of the blending of paper and electronic practices, highlighting the transformative potential of RDA, and the future encoding standard BIBFRAME (the Bibliographic Framework), which was then in development, to be a pair of linked data technologies that would allow library data to become interoperable with other existing metadata standards on the web. While this overall picture is positive, [Paper 3] illustrates certain individual data elements in the library catalogue, rooted in paper materials, struggling to be reconciled with prevailing information practices on the web. The analysis of a library user's frustration and success in [Paper 3] exposes the consequences of the incongruity and offers a framework to re-conceptualise certain aspects of cataloguing. [Paper 6] represents an application of this re-conceptualised framework to a specific data element—medium of performance—and shows the complex baggage with which a single data element can be laden.

Encoding standards for bibliographic records

In a traditional library catalogue, entries are considered surrogates of the physical bibliographic item. The concept of the bibliographic record is a relatively new one. In the *ALA Glossary of Library and Information Science*, the term "entries for a catalog" appeared through the 1943 edition, and was replaced in the following edition in 1983 by the term "bibliographic record" (Thompson, 1943; Young and Belanger, 1983). The emergence of the "record" in the intervening four decades reflects the widespread standardisation of not only the contents of the surrogates, but also the format in which these contents appear. The importance of the standardised format rose to prominence due to both the increased cooperation and sharing of catalogues between libraries, and to the computerisation of cataloguing in the 1960s. In this computer-assisted environment, computer and network technologies enabled cataloguers to transmit and store their work in a central, shared database, from which identical catalogues entries could be printed and put into service in multiple libraries. Thus, terms from database design such as record, format, fields, and data elements entered cataloguing vocabulary in this period.

MARC, the standard encoding format developed in the 1960s and still widely used today for records containing bibliographic description, mimics the cataloguing process on paper, with fields designed to accommodate the syntax of the content standard. As the content standard evolved to AACR2 and then to RDA, MARC continued to be revised to accommodate RDA data elements. Despite ongoing revisions which enhanced authority control and expressions of bibliographic relationships, the fundamental design remained.

In essence, this design of MARC encoding stores bibliographic descriptions while preserving the visual appearance of the catalogue entry (Griscom, 2022). In fact, one of the aims of computerising cataloguing was to facilitate production and use of catalogue cards. Early displays of electronic bibliographic records on a computer screen displayed the content and format of catalogue cards contents in exact replica. The design of MARC focuses on capturing text strings, sorting them, and re-arranging text segments for printing and on-screen display. This works well when the items are primarily a single type of materials, in this case, bibliographic. Electronic materials, however, possess a different set of characteristics. Therefore, as MARC continues to be the dominant bibliographic record format, the limited range of expression of the format becomes a constraint for the cataloguing of electronic materials. Nonetheless, today, the continued revisions of MARC to serve bibliographic description represents a second layer of retrofitting —- the retrofitting of the encoding format over the first layer of retrofitting of content standards.

[Paper 6] critiques recent efforts to revise the medium of performance field in the MARC data model. The syntax of the medium of performance field resulted in a method of recording data that would print the bibliographic description on paper or display on a screen adequately, but would fail to be indexed properly. The reason was that the content standard conflated several concepts whose distinctions are obvious only to a human reader. A single entry includes four pieces of data: the printed part (musical notation representing a musical layer of a piece of music), the instrument (for which the part is intended), players of the instrument, and, musicologically, the abstraction of a larger musical work that conveyed those previous three concepts. As a result, the semantics in the MARC format have become context dependent, unparsable algorithmically. The solution presented in [Paper 6] involved re-conceptualising this area of description and remodelling the informational content altogether. This type of work, taking an active role in the design of data models, might be considered beyond traditional cataloguing practice, but cataloguers have generally found current bibliocentric models being stretched too far, especially as more non-book-like concepts become more heavily relied on in library services involving non-book materials.

On the structural level, the design of MARC as key-value data structure reflects the operational needs and technological limitations of the 1960s: to store instances of mostly independent records with variable use of fields, and to be optimised for storage and simultaneous access. This choice of structure inadvertently reinforces the notion of independent records for bibliographic items, which is often not the appropriate structure for electronic materials. In the case study in [Paper 1], describing the relationship between the physical objects—photographic prints and plates—and their derivative digitised images required selecting a schema that could capture the hierarchical source-derivative relationship. The notion of the bibliographic record based on individual bibliographic items in traditional cataloguing fell short.

What could be done to maintain the relevance of cataloguing and the catalogue in a Blended Library? The Blended Library has so far compelled cataloguers to perform retrofitting of bibliographic description to electronic materials, and retrofitting of the encoding of bibliographic records to the retrofitted bibliographic description. The larger picture, however, points to the need for the traditional library catalogue as a whole to be retrofitted for it to be integrated with the contemporary information ecosystem. This presents an opening for library cataloguing to include the redesign of the library data operations to interact with other technology sectors. This interaction with RDF-style linked data technologies is explored in [Paper 2]. At that time, the call to move entirely away from MARC had also been raised by theorists and practitioners alike, such as in Coyle (2011) and McCallum (2012), who advocated replacing MARC with a new bibliographic format based on the linked data technologies. Today, this has materialised as the BIBFRAME standard. However, [Paper 2] differs in its perspective and outlook. Rather than suggesting an RDF-style linked data bibliographic format such as BIBFRAME and continuing to carry out traditional cataloguing in this new format, [Paper 2] advocates for a re-evaluation of the utility of the bibliographic record. It asks whether creating bibliographic surrogates would still be a fruitful approach to organise bibliographic description, whether cataloguing should be an activity limited to describing bibliographic characteristics, and how data originating from outside the library could be blended together to create descriptions not only for a bibliographic catalogue, but for the whole information resource in the whole information ecosystem.

Approaches to Cataloguing Electronic Resources

Electronic materials cataloguing

The life cycle of digital materials in a library follows a different workflow from that of physical materials. Printed materials come from a physical delivery. Each item is examined and catalogued by a cataloguer before “going into service” on the shelf. The role of the cataloguer is to create the surrogate in a record which becomes part of the catalogue. For most of the 20th century, this catalogue was printed out as cards and filed. In the 1990s, when electronic materials rapidly increased, coinciding with the end of large-scale catalogue card printing operations, the cataloguing community began to explore radically new ways to describe and organise these materials. These attempts have informed the Decomposition-Assembly Approach advocated in this critical commentary.

Early responses to cataloguing electronic materials

Early responses in the 1990s to cataloguing electronic library materials fell into two camps: one expanded the use of automation to extract textual information from electronic materials to populate MARC fields; the other performed parallel cataloguing using MARC and SGML (Standard Generalized Markup Language) markups, and then used automation techniques to embed the markups back into the electronic materials and converted the markups into MARC. Neither method has been widely adopted today, but both represented the need for expanding the scope of library cataloguing.

Automated MARC records creation

There was a period in the 1990s when cataloguers and cataloguing departments were optimistic about bibliographic control on electronic items and internet resources. Considerable efforts were put into early attempts at mapping, gathering, and automating the process of creating MARC records. For example, the computer program *On the MARC* described in Davis-Brown and Williamson (1996) is a sprawling macro that frees cataloguers from keying and re-keying data: the program extracts portions of an electronic text, adds ISBD punctuation, searches against LC Authority Files, then places the field data into a MARC record. Another example, the software *Alcuin's Little Helper*, described in Morgan (1996), was developed to create MARC records based on processing local document resource database outputs and extracting text from *Alcuin*, a database of resources on the world wide web.

Parallel SGML cataloguing

Others had noted the shortcomings of chapter 9 of AACR2, which analogises computer file cataloguing to book cataloguing (Brugger, 1996), but instead advocated for the use of the SGML family: Brugger (1996) mapped TEI (Text Encoding Initiative) headers to USMARC; Caplan and Guenther (1996) mapped Dublin Core to USMARC; Davis-Brown and Williamson (1996) marked up archival finding aids with EAD; McMillan (1996) used TEI markup for electronic theses and dissertations; Seaman (1995) embedded TEI headers in image files; Mandel and Wolven (1996) catalogued in SGML, programmatically converted to MARC, and loaded data into a MARC database, but also advocated for less-than-full cataloguing for digital / web resources while still "achiev[ing] an optimal level of name collocation, subject access, version control and genre identification." TEI, an encoding standard for representing text in digital form, is still in use today in digital humanities applications; converting finding aids into EAD has become standard archival practice. However, none of the other approaches took root in library cataloguing practice.

Data decomposition and assembly

In retrospect, these attempts at automating MARC record creation or using markups contributed to retrofitting the book cataloguing paradigm. However, two methodologies continue to be relevant for cataloguing the Blended Library: the use of automation in the extraction and assembly of data elements, and the two-way use of embedded data, extracted from the resources for use in the catalogue, and assembled and embedded back into the resources. The innovations described in [Paper 1] and [Paper 4] would take advantage of these methodologies. The approach developed in [Paper 1] was a blended approach: cataloguers with expertise in AACR2/MARC performed cataloguing manually, the data were then programmatically assembled into VRA Core (a SGML family schema). The approach in [Paper 4] moves away from the book cataloguing paradigm and suggests using information embedded natively within computer files.

The case study in [Paper 1] begins with an existing custom-built relational database for photographic prints and plates, some of which were being digitised. In the digitisation process, technicians and archivists would enter data into this database. This database worked well for a quick, local lookup of a small collection of several hundred, or even several thousand items. But it was not apparent how this database could be integrated into the main library's discovery interface. [Paper 1] developed a process that included extracting data from the source, mapping the source schema to traditional elements of descriptions as well as a target metadata schema, drawing the scope of resource description, including defining where to deviate from traditional cataloguing, and assembling the data into useful records. Ultimately, the records were assembled into the VRA Core standard. The data were also ready to be reassembled into MARC or any other schema, such as KML as demonstrated in Part 2 of [Paper 1].

[Paper 4] lays out the capability of computer files to self-describe. Rather than supporting the search for book-like characteristics, [Paper 4] advocates directly incorporating embedded metadata in these computer files, expanding the scope of resource description to meet the data natively embedded in the resources. These embedded metadata do not necessarily conform to the book cataloguing paradigm, and the methodology, like the one described in [Paper 1], involves acquiring and transforming data into a decomposed state, ready for assembly into other schemas and formats. This cataloguing approach helps create library data geared toward contemporary library service goals, such as the known item search based on incomplete information described in [Paper 3]. This will be further discussed in the next chapter.

Trajectory of cataloguing approaches

Together, [Paper 1], [Paper 4], [Paper 2], and [Paper 3] represent departure from the early responses to electronic resources cataloguing in a trajectory of transforming and expanding the scope of cataloguing from the book cataloguing paradigm. [Paper 1] suggests that cataloguers work with their existing skills and familiarity with the book cataloguing paradigm to extract resource description for later assembly; [Paper 4] introduces an alternative route for extracting descriptions from within the resource; [Paper 2] discusses future paths for content and encoding standards; [Paper 3] challenges the paradigm through an analysis of a troublesome area of description and lays out the path to use RDF linked data toward a more generalised resource description paradigm.

Adaptations and Innovations of Blended Cataloguing in the Post-Digital Revolution Information Ecosystem

In the context of traditional cataloguing described above, the tension between the bibliographic universe and the wider information ecosystem is apparent and inevitable. Further discussion of this tension and its ramifications on the present and future outlook on the cataloguing profession will be discussed in chapters 5 and 6. In the meantime, this section discusses specific blended practices from the prior publications.

The term “Blended Cataloguing” refers to a cataloguing operation situated in a Blended Library that operates in a mixture of paper and electronic environments. Blended Cataloguing represents an expansion of scope from the traditional bibliographic data to generalised metadata as libraries expand collections to include electronic mediums.

The challenges of cataloguing electronic mediums have been discussed amply in literature, such as the early attempts and experiments discussed earlier in this chapter. This section will discuss two specific challenges from earlier literature alongside some innovative approaches and theories presented in the prior publications of this dissertation.

Blended Cataloguing as an adaptation to electronic materials

Blended Cataloguing reflects a reactionary adaptation to the mixed environment in which paper and electronic library materials must coexist in the library catalogue, while the library catalogue itself also coexists in both paper and electronic forms. Seasoned traditional cataloguers are already experienced with a variety of material formats, as content standards for cataloguing have included non-book formats such as music materials, cartographic materials, microforms, moving images, manuscripts, and realia.

In practice, cataloguers regularly encounter items that require negotiating and blending cataloguing rules governing different material formats. An electronic typescript of a textual work printed out on paper (for example, Homer's *Iliad* typed into a word processor, saved as a computer file, then printed out) is simultaneously a monograph, a computer file, and a manuscript. When libraries began collecting digital materials either by accessioning born-digital resources or by digitising existing materials, cataloguers faced a different type of challenge.

Davis-Brown and Williamson (1996) make the distinction between natively digital materials, such as software programs, and digitally reformatted materials, such as text, still and moving images, and digitised audio. These distinctions by media type fall neatly into the institutional setting such as the Library of Congress, where administrative divisions align to media types (Geography and Map, Prints and Photographs, Motion Picture/Broadcasting, and Sound). However, confusions arise when digital copies of natively digital items are treated as reformatted and native digital audiovisual materials coexist with them.

Morgan (1996) further problematised the nature of materials: traditional library materials are physical and durable, whereas digital resources are ephemeral and fungible. Morgan (1996) recommends an automated periodic integrity check on digital resources, while Dillon and Jul (1996) suggest developing a protocol where the digital item and surrogates would regularly communicate and update each other. Both anticipated that a system of persistent identifiers would materialise (such as PURL and DOI) to provide a layer of abstraction, a location-dependent naming of electronic resources.

It became clear that electronic resources do not conform well to the traditional division of cataloguing by material type, or to the fundamental cataloguing concept of the surrogate. In fact, the ability of electronic resources to also embody every other material type calls into question the dichotomy between paper and electronic. Secondly, it also became clear that the bibliographic approach to cataloguing posed an impediment to working with electronic resources, and the needed layer of abstraction is one of generalised metadata that transcends material type and physical location.

Cataloguing as metadata curation

Milstead and Feldman (1999) was widely cited as the first paper to show the interweaving between library cataloguing and the rise of metadata creation in the information ecosystem as the internet developed and matured. While there was scepticism over the generalised metadata approach as argued in Gorman (1999) who called it a “false choice,” Howarth (2009) showed a clear confluence of bibliographic control and metadata applications. [Paper 2] focused on the convergence of standards as a driver toward expanding library cataloguing, and the potential of the RDA content standard and the BIBFRAME encoding standard to be made interoperable with other metadata standards in the wider metadata ecosystem.

At a more hands-on level, [Paper 1] applied the metadata approach to cataloguing photographs and their digitised images. This metadata approach represents the needed layer of generalised metadata abstraction. In the workflow developed in [Paper 1], although working with a MARC-like interface, cataloguers were, in effect, asked to create generalised metadata for the VRA Core format. This approach bypassed the bibliographic approach by breaking the synonymity between the bibliographic item and the book, decoupling cataloguing from bibliographic control and bibliographic records. The decoupling necessitated the development of a crosswalk, among the first of its kind presented in [Paper 1], between these bibliographic elements and the VRA Core, a schema designed for visual materials.

Following this approach, [Paper 4] expanded the scope of cataloguing into the self-describing features of computer files. [Paper 4] discusses accepting digital audio files as they are, in their own ecosystem. The role of cataloguing in this framework involves curating these embedded data, which can happen through developing means to extract them and designing the assembly into end user scenarios. Data curation activities turn the traditional cataloguing concept of following prescriptive rules on its head—-the practice of recording data elements according to prescribed content and encoding standards.

Cataloguing beyond the bibliographic record

In the 1990s, as cataloguers became increasingly aware of the limitations of the content standard for bibliographic description and the encoding standard for the bibliographic record, some responded by proposing the use of computer software for MARC record creation. Others responded by proposing the use of markup languages to encode bibliographic description. The prior publications offer another point of view that makes explicit what cataloguers have been doing all along, which is curating metadata. Generalising the cataloguing landscape enables situating the bibliographic universe in the universe of information resources. Conceptualising the bibliographic description is but one specific form of resource description and the bibliographic records are but one of the many possible outputs of cataloguing.

This new conceptualisation leaves the bibliographic record in a tenuous state, as the boundary demarcating bibliographic records is all but dissolved. Library catalogue entries are no longer bound within the four edges of catalogue cards or by the record terminator of MARC records. The generalised metadata framework described in [Paper 1], [Paper 4], [Paper 2], and [Paper 3] adds a necessary layer of abstraction. The focus of cataloguing shifts from producing bibliographic description to decomposing descriptions into individual data elements that can be later assembled into bibliographic description or any other forms of output. With this Decomposition-Assembly approach, the cataloguer sets data policy and curates library data to meet library service demands. The cataloguer makes judgement about boundaries—-what to include, what to exclude, judgement about standards—-and about which set of rules to follow while maintaining interoperability. Cataloguing now also includes the skills of assembling data elements into a variety of formats, schemas and presentation formats. This expansion of cataloguing will have a significant impact on the library catalogue. While traditional cataloguing standards have maintained the library catalogue in a static form, cataloguers can now adapt the underlying data to even the most nebulous service demands, such as the use case analysed in [Paper 3]. These transformations of the cataloguer, driven by the Blended Library, empower cataloguers not only to play an expanded role in the design of library services, but also to bring libraries into leading roles in the larger information ecosystem.

	Library Materials	**Library Services**	**Catalogue Presentation**	**Cataloguing Standards**	**Cataloguing Workflow**
Paper Library	Paper	On paper	In print	AACR2 / ISBD	Manual data entry and record assembly
Automated Library	Paper	On paper	In print + on computer screens	AACR2 / MARC	Manual data entry
Blended Library	Paper + Electronic	On paper + Electronic	In print + on computer screens	AACR2 / MARC + Spreadsheets + Dublin Core + a small number of metadata schemas	One at a time + Manual data entry aided by large-scale assembly by computer scripts/programs
Electronic Library	Electronic media	Electronic + Online	On computer screens	AACR2 / MARC + any schema + create new schemas	Above + programmatic harvesting of data

Table 2. Library Services and Cataloguing Characteristics in the Expanded Buckland Framework. This table summarises characteristics of the Paper Library, the Automated Library, and the Electronic Library as presented in Michael Buckland's *Redesigning Library Services: A Manifesto* (1992). Additional shaded cells are added to capture characteristics of the Blended Library.

Chapter 4 - The Catalogue Re-Imagined: the Decomposition-Assembly Approach

Introduction

The modern library is expected to collect a range of materials well beyond the capabilities of a catalogue that supports the organisation of primarily books. While treating microforms, maps, photographs, films, audio recordings and motion pictures as bibliographic items has been adequate, digital mediums present a challenge to this approach. Moreover, the same technologies that enable these digital mediums have also enabled the library catalogue itself to be computerised and to be transformed into an electronic information source. Technological developments that characterise the Digital Revolution have only accelerated the need to transform cataloguing and the library catalogue. However, cataloguing, which creates the catalogue, and the OPAC, which puts the catalogue in public service, have followed divergent paths.

Throughout the past few decades, new uses of the computerised, online library catalogue vastly outpaced cataloguing practices that support these new uses. Miksa (2012) presents a thematic overview of the history of the library catalogue and an analysis of the twenty-first-century legacy of the dictionary catalogue, the format which continues to be prominent today. On a more technical level, Leazer (1993) focuses on the challenges of the first three decades of computerisation of the library catalogue, from the 1960s to the 1990s, in the context of the theoretical lineage of Charles Cutter, Melvil Dewey, Seymour Lubetzky and Richard Smiraglia. Both authors reveal the catalogue's relationship with, as well as tension towards, evolving library collections and public service.

This chapter continues this investigation into the 21st century and focuses on the significant impact of the library catalogue as it becomes subsumed in the broader technological landscape, where cataloguing for bibliographic control has transformed into organising metadata curated in a generalised information ecosystem for use in library service. Specific use cases and examples are drawn from the prior publications to illustrate this transformation from the "transcribe-record" approach to cataloguing to the generalised "Decomposition-Assembly" approach to metadata organisation. Following sections in this chapter describe this change of approach resulting in the decoupling of the catalogued metadata from the bibliographic item, which enables a re-imagining of the library catalogue.

Approaches to Cataloguing: From Transcribe-Record to the Decomposition-Assembly

The library catalogue through computerisation

The library catalogue, traditionally defined, is the output of cataloguing, a technical operation among library functions. The form of the library catalogue developed, evolved, crystalised, and plateaued in the late nineteenth century into a dictionary catalogue, in which entries containing descriptive information of library materials are filed by titles, names, and subject areas in an alphabetical sequence. While the invention of the dictionary form of the library catalogue has often been attributed to Charles Cutter in the late nineteenth century, the chief promulgating factors, as summarised in Miksa (2012), were the combination of the widespread, low-cost distribution of catalogue cards by the Library of Congress Catalog Distribution Service beginning in 1902 and the codification of cataloguing activities in subsequent decades surrounding the practices of the Library of Congress. The effects of this history of coalescence around a cataloguing tradition of transcribing and recording data from bibliographic items into whole-item, single-subject, natural language entries are detailed in Miksa (2009). These features have since crystalised into what is now called the traditional library catalogue, the primary bibliographic service tool in the Paper Library stage of Buckland's evolutionary framework (1992), and the quintessential output that defines the cataloguing profession in the traditional sense.

Both Miksa and Buckland identified the computerisation of the library catalogue beginning in the 1960s to be the most disruptive event to the established cataloguing tradition and to the role of the library catalogue among library services. While Buckland, in his "Redefinition of the Library Catalogue" (1988), contemplated efficiency gains from catalogue automation, Miksa (2009) saw computerisation as a shift of technical burdens from cataloguers who create paper catalogues to system administrators who manage computer catalogue system, and, later, as a cause of significant increase of overall workload due to superimposing additional layers of vendor-automated discovery and retrieval systems on the library catalogue system. However, these two differing views arrive at the same conclusion: the scope of traditional cataloguing narrowed, as the dictionary form of the library catalogue remained in a perpetual stasis despite changes in supporting technologies. While Miksa (2009) reacted to this state of affairs with lamentation, Buckland (1992) described this state, the Automated Library, as a necessary stage of the evolution of library service toward the eventual fully dematerialised "Electronic Library."

The library catalogue in the internet era

The 2010s saw another wave of technological convergence. [Paper 2] describes this event as one that propelled a drive toward integrating library catalogue data as part of a larger information ecosystem. During this period, a new set of cataloguing rules (RDA) was released and a new bibliographic data format (BIBFRAME) was planned to meet the maturing linked data technologies whose tools were becoming popular beyond computing specialists.

Anticipating the impact of these converging technologies on the library catalogue, Buckland (1988), synthesising earlier technical and theoretical literature (Shera, 1972; Kochen, 1972), pondered a reunion with the activity of bibliography. In his view, it had been artificially bifurcated from information services and ceded to non-librarians.

Regarding library cataloguing, Miksa (2012) continued to advocate decoupling the activity from the product without abandoning the traditional library catalogue. He asks, "...could a solution be devised that could at least link to that past without adopting its rationale?" (Miksa 2012, p. 24) Written independently around the same time, the analysis in [Paper 2] addresses this question by imagining cataloguing as a dialogue in a Blended Library environment between the traditional library catalogue and the larger information ecosystem, with the cataloguer's role being one of connection and curation. To perform this new role, a new approach to cataloguing is needed.

The remainder of this chapter focuses on discussions and demonstrations of Decomposition-Assembly, drawn from the prior publications, as a key approach to the catalogue of the Blended Library, and as a means to achieve the prognostications of Buckland and Miksa.

The library catalogue in the information ecosystem

In his critique of new editions of two major cataloguing textbooks, Miksa (2009) writes, “the new digital realm does not simply represent a new kind of resource to be integrated into the given systems but rather a new approach to the value and use of information that is not even on the same page as the given systems. In this light, it seems appropriate to say that if ever there was a place to discuss library cataloging in relationship to them, it would appear to be in comprehensive texts of these kinds.” Referring to the content standard of traditional cataloguing (AACR2) and the encoding standard promulgated by the Library of Congress (MARC), Miksa juxtaposed “the given systems” on which the instructional content of the textbooks was based with the fast-proliferating web-based information ecosystem at the time. He continues to contemplate whether, for these textbooks, “a chapter [would] be warranted on how to create catalogs and classificatory systems from scratch that are not clones of the present given system. Such a chapter would need to include such things as how to determine catalog system objectives, how to determine appropriate data for information resource representation, and how to organize and display the data and, perhaps even more important, show how these basic matters are interrelated.” [Paper 2] describes this very interrelation centred around library policies on linked data technologies: that catalogue systems would be integrated into a web-based linked data information ecosystem, that the determination of appropriate data had been updated to the linked data compatible to the RDA standard, and that the organisation and display of data would follow linked data structures such as BIBFRAME. [Paper 2] continues to anticipate that cataloguers in such a technology environment would begin to engage in the type of work Miksa (2009) described.

Buckland had anticipated the move toward dematerialisation of libraries in general. But at the level of the library catalogue, Miksa suggested that it should transform alongside, but without describing, in practical terms, the types of skills and training that would be involved to curate, manage, and present data. The research and demonstrations in [Paper 1], [Paper 3] and [Paper 4] would suggest that, even before later textbooks editions added chapters on linked data, cataloguers were already manipulating and managing library data beyond the traditional confines of

AACR2/MARC into bridging library data with other metadata schemas, finding their way in this generalised metadata ecosystem.

Taken together, the work described in these prior publications points to a distinctive approach particularly suited for libraries. In presenting information to library users, libraries generally do not offer pre-selected, packaged information products, but rather provide a custom-assembled, tailored response to each user enquiry, or to a particular collection or service that the library may happen to own or choose to maintain as a starting point. On the supply side, library catalogues, as incompatible with the information ecosystem as they may be, are a significant high-quality source, alongside a plethora of non-library sources that each has its own characteristic, often with a non-interoperable data structure. Given the nature of the data flow, decomposition is required in order to manipulate the source data so they can be reassembled into the desired form of output for library services. This Decomposition-Assembly Approach is familiar to computer science, and, coincidentally, is shown to be a unifying factor in the transformation of the work of the cataloguer.

The following sections of this chapter further detail the contributions of the prior publications as this approach made clear that the scope of the cataloguer's work would go beyond producing and toward transforming the form of the library catalogue.

The Catalogue in the Linked Data Superstructure

[Paper 2] situates the library catalogue in a critical juncture where the data technologies, bibliographic model, and cataloguing rules were converging to revolutionise the library catalogue. These changes were vast and sweeping, affecting not only the structure and contents in the library catalogue, but also workflow, scope, and conceptualisation of cataloguing.

Library cataloguing encounters linked data in heterogeneous data sources

One of these changes was the transition of the content standard for cataloguing from AACR2 to RDA. First released in 2010, the full implementation of RDA by the Library of Congress in 2013 marked a significant departure from its predecessor. In addition to building on the Resource Description Framework (RDF), an entity-relationship model compatible with linked data technologies, RDA also offered wider latitude to the sources of information to be incorporated into bibliographic description. In particular, for computer files, RDA rules now allowed information to be sourced from embedded metadata, which prompted the investigations in [Paper 1] and in [Paper 4]. This rule change freed cataloguers from the bibliographic universe to information objects of all kinds, especially electronic ones, which could now be described natively and more efficiently. The challenge of using embedded metadata in bibliographic description is that these data, unlike printed text, are not meant for the eyes of the cataloguer, but rather to be processed by computer programs. When building catalogues in this manner, it became necessary to extract data from one source and place them into another. This approach is especially useful for handling non-bibliographic data. In essence, the Decomposition-Assembly becomes the necessary, de-facto approach.

The Decomposition-Assembly Approach

In the general sense of the terms, decomposition is the separation of an entity into its constituent units, whereas assembly creates a new entity by combining these smaller pieces. Distilling from the library metadata use cases illustrated in the prior publications, the Decomposition-Assembly Approach describes the process through which characteristics and properties of materials and information are decomposed, and then assembled into a form that suits the proposed or intended use. The decomposition process results in ontology-agnostic RDF triples which serve as the basis for generating library metadata, including library catalogues, whereas the assembly process involves automated machine data processing which combines units of interoperable data elements that are preferably atomic, that is, decomposed to the greatest extent possible. This approach reflects a direct implementation of RDF, capable of reconstituting RDA data elements in the MARC format. Simultaneously, it offers a cataloguing strategy focused on adaptive design and flexible handling of data elements. This positions the library in a more attendant state of readiness to meet any demands created by materials, institution, and users, and, consequently, preserving long-term utility of the data.

In an ecosystem where metadata are created by the Decomposition-Assembly Approach, the cataloguer participates in a cycle of curation, policymaking, and creation (Figure 1). This participation can be seen as a practical approach for the cataloguer’s transition into the role of the librarian as exposited in the “Theory of Metadata Enriching and Filtering” by Alemu and Stevens (2015). Bridging the parlance in their theory which “requires a shift in the role of the librarians from metadata creators to metadata systems architects…experts at providing structure, granularity, and interoperability…” (p. 99-100), structure is created by the assembly process; granularity and interoperability are created by the RDF-style decomposition process. The cataloguer’s policymaking activities involve design and development of the assembly target, that is, standards and platforms that enable the library to become what Alemu and Stevens called a knowledge-building environment which supports the social construction and filtering of metadata. While Alemu and Stevens emphasised the significance of library users as co-creators, their model did not consider librarians themselves as users, especially in the course of providing information service partnering as “co-co-creators.” The metadata, whether created by the cataloguer (“*a priori*”) or a result from the librarian-user co-creatorship (“*post hoc*”), become available to curatorial activities and decomposition, regardless of origin, so long as the metadata are appropriate and suitable. The emergent areas of the cataloguer as data curator, policymaker, and knowledge creator are further detailed in Chapter 6.

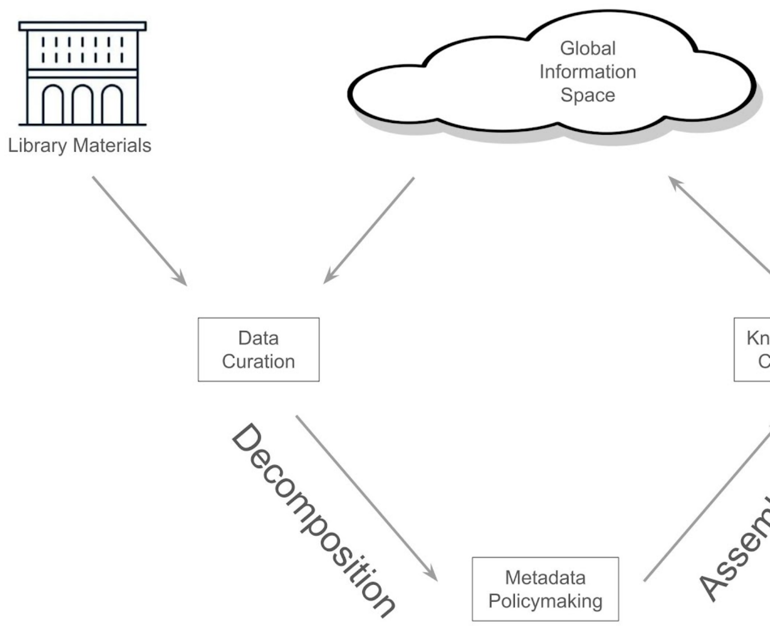


Figure 1. **The Decomposition-Assembly Approach**, as followed by the Blended Cataloguer. Data are curated from library materials and from the global information space and decomposed. With library user participation, the target metadata structure is designed and developed through the policymaking process and becomes standardised. The decomposed data are then assembled accordingly, resulting in newly created knowledge.

Use cases of Decomposition-Assembly

[Paper 1] described a novel workflow where manually-input MARC-style data were decomposed and then, combining with data from other sources, assembled into a non-bibliographic catalogue in the VRA Core and KML schemas. This use case is particularly relevant to describing non-book items for which non-bibliographic schemas are more often used. For data already in the traditional library catalogue, [Paper 3] illustrated the urgency to re-examine certain areas of description, such as medium of performance, so they could be decomposed and reassembled in linked data structures. [Paper 3] also illustrated that this re-modelling work is well suited to the existing expertise of library cataloguers.

This urgency, more broadly, of employing the Decomposition-Assembly Approach is further illustrated by the bibliographic search scenario described in [Paper 3]. The library users did not take the “traditional” route that would have involved waiting a few hours or a day for the opportunity to make a trip to the library, a conversation with a librarian, a look in the library catalogue at several records, a walk to the shelves, and opening up a few volumes to look at the table of contents and flipping to the pages. Instead, the users performed a mixture of activities: queried internet search engines, looked at crowd-sourced indexes on Wikipedia, perused tables of contents in physical volumes, and, with the assistance of some prediction / suggestion algorithm and a stroke of luck, arrived at the answer. While the outcome was successful, the scenario revealed that relevant information was scattered, and the route to attaining the information was labyrinthine. The questions then raised in [Paper 3] were how the library catalogue could be used more fully, and what could be done to connect the catalogue and disparate external data sources into a unified presentation.

With electronic resources becoming increasingly accessible and prevalent, one expedient path is to harness their embedded metadata. [Paper 4] presents a comprehensive overview of their foundational metadata structures for multimedia materials. Commercial software (such as iTunes, which uses embedded metadata and CDDB) and other non-library initiatives (such as CDDB/Gracenote, freedb, MusicBrainz, etc.) have developed organisational tools, but have struggled with a standardised method of decomposing data for complex names and titles, an area of practice highly familiar yet under explored by cataloguers. For textual materials, [Paper 5] takes this work beyond using existing data standards, but to develop one, in this case, electronic music scores, in a way that facilitates interfacing with data in library systems.

These use cases of the Decomposition-Assembly approach combine bibliographic data with data external to library systems. With this approach, the assembled output does not necessarily need to resemble the library catalogue, which is traditionally a static, textual document with a tabular presentation. Rather, it can be re-imagined as a dynamic, multi-format knowledge source.

The library catalogue re-imagined

When data become available outside the MARC structure, cataloguers become free to assemble them in a variety of ways, thus affording new forms and functions for user service. Non-library data can be searched and displayed seamlessly alongside library data, supplementing bibliographic data with data derived from external sources and the electronic materials themselves: embedded descriptive and technical metadata, as well as extracted metadata describing the materials' contents. Such an augmented library "catalogue" would not only display static information, but would be displayed in a variety of contexts and provide embedded analytical functions, which are potentially more helpful and insightful to the user.

Library data as annotation

[Paper 1] illustrated one of the earliest examples of a cartographic presentation of a library catalogue where catalogue descriptions of a photographic collection were superimposed on a web map service. The underlying data, originally in an amalgam of AACR2/MARC and VRA Core formats, contained externally imported geographic metadata (city and longitude/latitude coordinates), and was assembled into the cartographic KML format for use in Google Maps. The Google Maps display interface, through a combination of built-in capability and local custom programming (in PHP), performs calculations to generate pegs suitable for each location, combine entries into a single peg that are in close proximity, and overlays the information on the default road map format offered by the Google Maps service. [Paper 1] further discussed overlaying the data on digitised historic maps, an implementation that was not possible with the Google Maps service at the time but became widespread a few years later. Today, a variety of presentation formats on commercial web services have added the capability to insert annotations, such as specific location and time on cartographic maps, on particular words of characters in serial texts (chats), and at specific time and space in audio/video. For library data to be relevant in these settings, a re-conceptualised library catalogue would require fully decomposed source data to afford the flexibility for being reassembled and inserted.

Library data supporting auto-generated catalogues

[Paper 4] describes how embedded data are foundational to self-describing electronic resources that allowed software to create auto-generated catalogues of local collections on personal computers. The more generalised concept of providing self-generating catalogues based on embedded metadata in electronic materials is not uncommon, as seen in many consumer-level software that manage PDF files, image files, audio files, and video files. However, the lack of vigorous standards has hampered the usefulness of these management systems. [Paper 5] highlights the significance of participation by libraries, especially cataloguers, in developing standards in these material formats. In the functional requirements proposed in [Paper 5] for electronic music scores, not only will the score provide traditional bibliographic catalogue information (title, author, etc.), but also actionable metadata concerning the musical contents. This enables self-generation of a catalogue that is useful not only as a bibliographic catalogue, but also as a catalogue that can be searched by musical characteristics (such as key, vocal range, etc.).

Library data supporting analytical processing

An analytical catalogue environment augments the static catalogue by offering underlying data that support dynamic processing and calculations. The mouse-over display of descriptions, thumbnails, and hyperlinks described in [Paper 1] was an early demonstration of an augmented display. Subsequently, [Paper 5] would expand this concept to embedding metadata into the electronic music scores in such a way so cloud applications could do more than search and retrieve; they could perform analytical processes within the user environment where musical contents could be manipulated dynamically prior to retrieval (such as transposition, re-instrumentation, etc.).

The search scenario in [Paper 3] is much more complex. Besides the ability to gather scattered information from a number of sources, the success of the search required overcoming several misremembered and misspelt terms, and hinged on an automated suggestion. What [Paper 3] advocated was a decomposing of library data such that the data could be available, in real time, to search algorithms. Whether pre-programmed or achieved through machine-learning, search algorithms cover a large swath of the information ecosystem, which, in turn, enables answering generalised queries.

Zoomable catalogues

By combining these new forms and functions, the catalogue will achieve another characteristic that enables navigation between entities as well as between levels of granularity. For example, including musical content metadata in the functional requirements for electronic music scores simultaneously enables editing and manipulation of the content as well as supplying descriptions that can be processed at the level of the document, section, voice, melody, bar, figure, and individual musical event. As described in [Paper 5], users are able to search, edit, compare, and navigate to any level and move between them. [Paper 6] described a similar ability for navigating levels of medium of performance through the taxonomy of musical instruments, characteristics of performers, and sections and subgroups within ensembles and orchestras.

This characteristic of “zoomability” had been theorised in Kochen (1972) to describe a contextually hyperlinked electronic text environment where users had the ability to zoom in and out. Users could navigate the text from hierarchical, successively detailed outlines and jump from place to place within the resource, locate concordances (i.e. word/phrase search within text), and zoom from resource to resource—that is, navigate away to another related text. Since that time, zoomability has been realised in web documents through hypertext linking in texts and maps. For example, a PDF file can display a resizable image containing text, together with an invisible layer of encoded text corresponding to the exact location of the text within the image; map services on the web can contain multiple data layers pegged to geographic coordinates, as demonstrated in [Paper 1], based on the level of viewing detail. The pioneering technologies for rendering zoomable displays in other electronic formats such as audio, video, and music scores as described in [Paper 3], [Paper 5] and [Paper 6] have been used today in specialised library applications such as those described by Lewis *et al*. (2022) and Lewis, Page and Dreyfus (2021), but the practice is not yet widespread.

Zoomable displays allow multiple data sources to be juxtaposed and combined. New insights come through this process—the sense of proximity and distance displayed between images (described in [Paper 1]); discovery of new musical instrumentation groupings (the model proposed in [Paper 6]); new methods of musical analysis performed on electronic music scores (a scenario in [Paper 5]). These insights are unavailable from a catalogue-style display. However, libraries have yet to adopt the production of zoomable displays as part of their services. Should cataloguers adopt the Decomposition-Assembly Approach, the ingredients and workflows would be in place for libraries to become major players in this area of knowledge production.

The zoomable catalogue as bibliography

Incidentally, in the same year Kochen described “zoomability,” Shera (1972) has remarked that bibliographies, the products of the intellectual work of indexing the existence of works and titles by discipline or topic, have been spun off from library operations to commercial vendors, while cataloguing, which is primarily concerned with physical library items, has remained in-house. Spurred by the Decomposition-Assembly Approach to cataloguing—a particular evolutionary path of the library catalogue towards acquiring annotatability, analysability, and zoomability—re-establishes the continuum between the two practices. As data in both can be dynamically decomposed and re-assembled, the book catalogue and the bibliography can be reunited in a single bibliographic environment.

The integrated online information environment enabled by the Decomposition-Assembly Approach to cataloguing might well take library service full circle, allowing it to reclaim the crucial area of bibliography, while at the same time regaining ownership and control of computer systems for cataloguing, display, and visualisation.

Re-imagined Cataloguing

What will the future library catalogue look like? The traditional catalogue has already become a subset of federated search engines and discovery systems that query an assortment of systems for metadata and deliver contents to users. This traditional library catalogue will continue to be a service tool for identification, collocation, and evaluation, but materials involved are now a blend of books and non-books, physical items and electronic resources.

The Decomposition-Assembly Approach has been shown in the prior publications to amplify the impact of the library catalogue as it coexists in the generalised information environment with non-library data curated by the cataloguer. Focusing on decomposing the bibliographic unit into atomic data elements, this approach opens up the flexibility to re-assemble the data toward novel uses. This re-imagined cataloguing, decoupled from the bibliographic unit, will be able to move freely between material formats and mediums and between levels of granularity. The result is a continuum between metadata and data actionable for processing and analysis, a re-imagined catalogue that integrates information seeking and knowledge production. Cataloguing will move away from being reactive, performed only when the cataloguer encounters a book. Rather, cataloguing will be a prospective, knowledge-producing activity that anticipates the dynamism of the general information landscape.

Chapter 5. The Blended Cataloguer: The Expansionary Trajectory of the Library Cataloguing Profession

Introduction

Library cataloguing has been recognised as a profession with a distinct skill set among library operations (Bair, 2005). As a working definition, profession is here defined to be a group of workers whose jobs require specialised knowledge from education, training, who work autonomously within their areas of expertise, and whose performance are judged against professional standards. This three-part framework of (standards, knowledge, and expertise) will be the basis of the discussion in the middle section of this chapter.

Chapter 3 and Chapter 4 have shown that the materials and the catalogue of the Blended Library present a major force behind the transformation of the professional nature of cataloguing work. This chapter discusses the professional aspects of cataloguing presented in the prior publications that together characterise the trajectory leading to the Blended Cataloguer.

This chapter's discussion will be presented in three parts. First, a summary of the trajectory is presented in two aspects: reactive and proactive. The reactive summarises discussions in Chapter 3 and Chapter 4, while the proactive previews Chapter 5 and Chapter 6. Then, exemplary professional elements are highlighted from the prior publications against literature on the historical evolution of the cataloguing profession. In the third part, discussion contrasts the "Blended Cataloguer," a term developed in this critical commentary, to the "Blended Librarian," a term coined by Bell and Shank (2004).

The Expansionary Trajectory of Blended Cataloguing

Chapter 3 and Chapter 4 presented cataloguing practice adapting to the Blended Library. These reactive adaptations to new library materials and new uses of the library catalogue also opened up areas of work not historically considered within the scope of library cataloguing. As shown in the prior publications, these new areas are adjacent to traditional cataloguing work and present natural expansions of its scope and the potential to shape future professional identity.

Adapting to the Blended Library: reactive aspects

The following summarises the adaptations presented in Chapter 3 and Chapter 4:

- Cataloguing blended materials: bibliographic control of both digital and non-digital resources
- Cataloguing for the blended catalogue: cataloguing for library services in the web environment
- The Decomposition-Assembly Approach: The manipulation of data and metadata for the in order to accomplish the first two. This approach is also the basis that leads to the proactive aspects in the expansionary trajectory.

Defining the cataloguer in the Blended Library: proactive aspects of the expansionary trajectory

Cataloguing in the perpetual Blended Library is characterised by expanding the scope and defining new roles of the profession proactively. The prior publications highlighted three roles of the Blended Cataloguer:

- Cataloguer as data curator: management of both data and metadata towards an organised, purposeful presentation as library service
- Cataloguer as knowledge creator: creation of an integrated information environment where the curation facilitates production of knowledge
- Cataloguer as policymaker: participation in standard setting developments in areas such as data modelling, data encoding, file formats, data transmission standards, and digital object functionalities.

These three roles will be touched on later in this chapter, and more fully discussed in the next chapter.

Expansion of the Cataloguing Profession

Introduction

Literature often approaches the history and evolution of the library cataloguing profession by examining job requirements in the field and skill sets included in LIS curriculum, training, and competency standards. A brief history of the roles and responsibilities of the library cataloguer is included in Han and Hswe (2010), as well as in Turner (2020) and Ivey (2009). The analysis in Han and Hswe (2010) grouped the roles of responsibilities of cataloguing librarians into three historical periods: (1) a relatively stable "traditional environment" through the mid-1990s; (2) the period of the "electronic environment" beginning in the 1990s, when cataloguing activities became increasingly dependent on automated cataloguing systems requiring technological skills and knowledge surrounding the MARC format; (3) the transition of the catalogue librarian to the metadata librarian since the 2000s, when increasing varieties of systems and resource formats were introduced and overlapped significantly with traditional cataloguing skills.

The cataloguing profession in Buckland's framework

In the context of the evolutionary framework of library services in Buckland (1992), the Paper Library and the Automated Library are analogous to the first and second periods of Han and Hswe (2010). A cataloguing profession solely consisting of metadata work on electronic materials would correspond to Buckland's Electronic Library. However, the Blended Library proposed in Chapter 3 more accurately matches the cataloguing profession according to Han and Hswe (2010)'s third and current period, in which traditional cataloguing coexists with cataloguing of a variety of material formats. This coexistence creates a situation where the input and output are both heterogeneous. Cataloguing practices have adapted to this, as shown in Chapter 3 (regarding the input materials) and in Chapter 4 (regarding the output).

Transition of the library cataloguing profession in the Blended Library environment

[Paper 2] described the moment when the cataloguing profession developed RDA, simultaneously a content standard for traditional cataloguing and a linked data standard capable of bridging traditional bibliographic description with metadata work in the linked data environment. The paper called for future cataloguing to "be done on the Web, for the Web."

The skill sets required for this expanded scope of work involves more generalised knowledge and tools that create and manage metadata through their life cycle, regardless of the material format and retrieval medium. These processes are presented and analysed in [Paper 1] for archival and digital images, [Paper 4] for digital audio, and [Paper 5] for electronic music scores.

The implications for the cataloguing profession are two: cataloguing will expand into a variety of materials regardless of the library's ownership, and cataloguing will turn its focus toward connecting disparate materials for more comprehensible and comprehensive presentations to library users.

Cataloguing becoming metadata work

Turner (2020) aptly described a situation in which, when the term “metadata” proliferated in describing cataloguing librarianship, the meaning and usage of the term became “chaotic” as institutions grappled with conceptualising the roles and skill sets required. In the decade between 2000 and 2010, Turner (2020) noted a period of rapid transition, with descriptions of cataloguing positions overlapping with metadata positions. And, as of 2019, job descriptions for positions whose titles included “cataloguing,” “metadata” and “cataloguing and metadata” often conflate these skills. Turner (2020) continues to note that while the cataloguing skill set continues to be integral to the profession, the scope of the profession has broadened to non-MARC areas.

Going back two decades, Milstead and Feldman (1999) described the “chaotic atmosphere of clashing standards” in bibliographic control of digital materials. Yet another two decades earlier, in his review of the initial publication of AACR2, Gorman noted that “The MARC record remains an automated version of a manual catalogue entry” (1978, p. 210) [Paper 2] showed the idea that equating library cataloguing with bibliographic record creation contributed to a sense of confusion and professional precarity, as well as the pointing to a need to disentangle the two and recognise the metadata work in library cataloguing.

Standards: expansion and bridge building

The greatest challenge of broadening standards used for cataloguing is expanding cataloguing’s purview of into non-MARC encoding formats and schemas. As Gorman (1978) described it, since MARC serves as a bibliographic record entry form, using non-MARC schemas means moving beyond the traditional approach of bibliographic record entry towards the Decomposition-Assembly approach. Several of the prior publications present case studies of how this transition unfolded in practice.

[Paper 1] shows an early example of this transition. The crosswalk between AACR2 and VRA Core 4.0 enabled traditional cataloguers to use a very similar workflow to enter metadata, yielding an end result where the data could be combined with other embedded metadata from the digital images.

Another example of this work involves deriving and extracting metadata that originate from audiovisual materials in their native formats. [Paper 4] anticipates this type of horizontal bridging from the vertically integrated metadata embedded in digital audio formats. This work "necessitate[s] mapping KLV-style metadata to XML" ([Paper 4], p. 256), where the knowledge of both the KLV and XML metadata models are new to traditional cataloguing.

(Note: KLV is the prevalent encoding format for metadata in digital audio as well as in digital images. It stands for the "key-length-value" data encoding method where each data field first declares its key and its field length, followed by the value of the field. The bibliographic encoding format MARC follows a similar concept but uses a "directory" where the key and length are followed by a pointer to another location in the record where the value of the field is stored.)

For manipulating embedded KLV-style metadata, the encyclopaedic exposition of the embedded metadata schemas discussed in [Paper 4] provided cataloguers the expanded technical details. In regards to using the XML format for the output destination, [Paper 3] detailed the conceptual structure of the entity-relationship model. Discussion unfolds from how one could conceptualise linked data from a reading of a spreadsheet—how a spreadsheet presentation of fields and values can be construed as entity-relationship linked data—to constructing linked data triples, where medium of performance data are used as examples. [Paper 6] then developed a fully built-out linked data model for medium of performance data that enables more nuanced description. Together, these papers showcase a bridge between traditional cataloguing and the metadata domain.

Knowledge: education and training

Cataloguing and classification have long been core skills taught in the LIS curriculum. While Salaba (2020) has found that information organisation is almost universal in LIS curricula in the United States, Alajmi and Rehman (2016) have found that traditional cataloguing is increasingly supplemented by metadata, ontology, and linked data technology topics in these courses. However, Joudrey and McGinnis (2014) have found that the sheer amount of content that needs to be covered is difficult to pack into a single course.

Over a decade earlier, Velluci (1997) advocated a “syndetic” structure of cataloguing instruction, where cataloguing concepts would permeate the whole curriculum. Later, Bawden (2007) advocated the teaching of information organisation and information retrieval in close combination, and Dobreski, Ridenour and Yang (2022) have found that that has indeed been the trend. The merging of organisation and retrieval topics reflects the proximity of cataloguing to the public service functions of the catalogue, and also highlights how the scope of cataloguing is encompassing both input and output. Implied is the need to include decomposition as well as assembly in the skill set.

Expertise: professional autonomy

The “assembly” aspect of the Decomposition-Assembly Approach involves considering public service, since the assembled output is what library users interact with. Traditionally, the development of search and retrieval systems was the domain of computer programmers in software companies. The need for specialised discovery systems for libraries arose, in part, due to the particulars of bibliographic cataloguing that are opaque and cryptic to machine processing. While integrated library systems and discovery systems are essential to library service, through data modelling work, cataloguers could affect the design of the structures of metadata fed into these systems.

Addressing the design of medium of performance model, [Paper 6] began the modelling based on user behaviours as analysed in [Paper 3], and the extensibility of the modelling was made primarily to accommodate retrieval. As analysed in [Paper 3], traditional cataloguing practices have long produced library catalogues for human users who could learn to understand and use them. Yet, with the semantics of the traditional structure, information could be lost when read by computers. The analysis in [Paper 6] worked to decompose data elements in medium performance and assembled them in an extensible, computing-friendly way, independent of their human context. Adding data modelling to the cataloguer’s skill set enhances machine processing, and the cataloguer also gains finer control of the data. The added capability to assemble data in non-proprietary, open standards reduces library’s dependency on software developers and increases autonomy for the cataloguing profession.

Gaining finer control of metadata also helps break the material format barrier. Decomposition enables freeing data from its carrier, whether embedded data from digital files as shown in [Paper 4], or traditional cataloguing data from the bibliographic record as shown in [Paper 1]. Its application enables "re-contextualization, repackaging, and commodification" of the materials ([Paper 4], page 256).

This type of re-integration of metadata is a natural extension of the trend to combine the study of organisation and retrieval, to see information production and consumption as a totality, and to be able to see information products through their entire life cycles. But for the library to participate in the ecosystem as a producer of new knowledge and new cultural products, it will involve more than skills and tools. Discussions on the policymaking aspect will appear in the following chapter.

The Blended Cataloguer and the Blended Librarian

In 2004, Steven Bell and John Shank coined the term "Blended Librarian" to describe the new roles assumed by public service academic librarians in the early twenty-first century. A parallel development took place for cataloguing librarians in the same period spanned by the prior publications. This section juxtaposes the concepts of the Blended Librarian and the Blended Cataloguer to contrast the responses by these two related professions to rapid technological change.

The internet radically altered the information ecosystem which resulted in significant changes to operations and services in libraries. That is a cause for reflection on the librarian profession as a whole. Buckland's framework suggests that ubiquity of electronic resources would ultimately propel the Automated Library to the Electronic Library. The previous chapters have shown the persistence of the period of transition and named it the Blended Library. Bell and Shank cast the Blended Library as a period of "tumultuous change" in the academic enterprise, which created a "critical professional juncture" where the academic librarianship profession was "struggling with ways to harness and weave new technologies into our existing fabric of high-quality information service delivery" (p. 372). For the cataloguing profession, the Blended Library represents a period of opportunity and radical expansion of the cataloguing profession. The struggle did take place, but earlier, roughly corresponding to Buckland's Automatic Library period. Chapter 3 and Chapter 4 show that, from the period from the 1960s through the 2000s, cataloguing practices adapted to computer automation through updating the standards and formats as libraries collected new materials and offered them through new avenues opened by new technologies. The new demands on cataloguing resulted in the expansion of knowledge, standards, and expertise summarised earlier in this chapter.

These forces of change that compelled the cataloguing profession to adapt raised alarm in some quarters: "MARC Must Die!" (Tennant, 1995); "Is MARC Dead?" (Coyle, 2000); "MARC Must Die! 15 Years On" (Tennant, 2017); "Data Catalogs are Dead; Long Live Data Discovery" (Moses and Saha, 2022). These sentiments were well summarised in Danskin (2006). Danskin described the list of "challenges facing cataloguing" as "daunting," citing the reluctance of institutions to cover the increasing cost of organising an ever-expanding information ecosystem and competition from other information sectors. Within five years, at the start of the period spanned by the prior publications in this dissertation, Cerbo in (2011) asked, "The debate over the need for catalogers, cataloging principles and set rules in this new 'Google age' of key word [sic] searching and web browsing brings forth the real question of the necessity of cataloging. Is there a need for library catalogers that there once was?" (p. 324).

Bell and Shank (2004)'s Blended Librarian also has existential anxiety from the loss of control of multiple aspects of library operations to other information sectors. These anxieties can be summarised into the three categories below. In each category, I draw parallels to the cataloguing profession and discuss areas where the issues have been addressed or could potentially be addressed in the future.

Commercial vendors creating self-siloed systems

Bell and Shank (2004) worried about courseware systems cordoning off faculty-student information needs and exchange from public service librarians, and textbook publishers shutting out the library by incorporating contents into textbook companion websites. In contrast, commercial forces have taken over library systems design while reinforcing traditional cataloguing. ILS system vendors created ever-sprawling management software that preserved traditional cataloguing practice, and other companies began to develop specialised format-specific software for storage, discovery and delivery of specific formats such as images, videos, and archival materials. The fragmentation of bibliographic systems created high demands for cataloguers to bridge these systems.

The pilot project presented in [Paper 1] demonstrated the bridge between traditional cataloguing and the digital asset management system that would eventually be built (VTLS VITAL). Metadata were collected through traditional cataloguing methods, but metadata was created and embedded to new standards (VRA Core 4.0 and EAD). At the time, the technology was new and the software did not even exist yet! To hedge against the failure of the vendor to deliver the software and against catastrophic system failure, extensive thought and effort was given to treating the metadata in a way that retained the ability to store and retrieve the metadata independent of the digital management asset system. In doing so, libraries gained more autonomy over their data and became more free to use or develop other systems that address user needs. While the primary interface of the digital asset management system offered a traditional OPAC search, the cartographic interface presented in [Paper 1] is a demonstration of what could be enabled by cataloguers' new-found skill sets.

Publishing ecosystem shutting out the library

Scholarly publishing models gradually moved away from journal publishing in library databases, at the same time publishers promoted personal subscriptions to content delivery rather than obtaining access through libraries. [Paper 5] addresses this issue in the area of electronic music scores where libraries extended library technical operations to encompass the life cycle of the digital object. Historically, as libraries expanded into collecting and offering digital materials, they have ceded the development of some of these material formats to other domains, most conspicuously electronic books and journals. [Paper 5] argues that the skill set of a cataloguer in manipulating embedded metadata could naturally be extended to include the design and standardisation of emerging material formats. The cataloguer could place libraries in a better position to provide the platform for the production process, the platform for consumption through discovery and access, and the archiving of the same materials.

Libraries now have the opportunity to become a partner in the development of digital formats for text, audiovisual materials, and geospatial applications. [Paper 4] reveals a vertical integrated production chain for audiovisual materials that focuses on the audiovisual content itself, while descriptive metadata inhabits a heterogeneous, somewhat chaotic landscape; [Paper 5] and [Paper 6] together provide a road map for expanding into the area of material format development and a data model that could be used in that format; [Paper 1] showcases a crossover combination of traditional cataloguing of photographs and metadata creation for retrieval and access.

Internet search engines bypassing the library

Bell and Shank (2004) are concerned about search engines such as Google, Amazon and Microsoft encouraging information seekers to bypass the library. The information seeking scenario described in [Paper 3] is a typical example. It is far from uncommon yet presages what libraries can facilitate in the future. More often than not, seekers begin their search with incomplete, inexact, and inaccurate information. [Paper 2] and [Paper 3] made the case for turning the bibliographic focus of library data from a constraint to a strength, if library data could become more complete and nuanced. This could be done directly through publishing library data to the internet so that the data are discoverable via general internet search engines, or through offering bibliographic data as corpuses for machine learning. Since bibliographic data have been siloed for quite some time, transforming the data into open linked data formats would require the cataloguing community a major effort to embrace Decomposition-Assembly as a definitive approach. [Paper 1], [Paper 2] and [Paper 3] have shown the potential for a significant volume of bibliographic data that could be converted, and [Paper 6] demonstrated the possibility of refining bibliographic metadata through remodelling. These data would be made more digestible to computing, whether in indexing, semantic parsing, or as a corpus for machine learning. [Paper 5] further envisions cataloguers participating in the entire metadata life cycle, as well as in the design of data models that facilitate information retrieval and other computing technologies.

Commercial search engines are here to stay. What libraries can accomplish, primarily with the expertise of cataloguers, is to bring libraries to a level playing field and become development partners. Bibliographic data, besides being rich and nuanced, have a bias toward accurate, non-commercial information. With the appropriate information interface, what libraries could offer would be a refreshing addition to today’s commercial forces.

Conclusion

In an analysis of cataloguing job titles and descriptions, Geckle and Nelson (2017) described the cataloguing profession as becoming more technical and sophisticated. Yet our job descriptions are “out of sync with our immediate future” and the profession “needs to figure out how to properly express what we are doing now and what we need to be doing in the near future” (p. 63). The authors aptly asked: “Who are we? What do we do? What will we be doing?” (p. 62)

The first question was addressed in Chapter 5, which discussed the distinct skill sets that define the library cataloguer. The second question was addressed in Chapter 3 and 4, which discussed the various elements of cataloguing work as the scope of library collections and library services evolved. The last question will be addressed in the next chapter. While Chapter 5 focused on the expansionary trajectory experienced within the cataloguing profession, the next chapter focuses on the outward trajectory, highlighting emergent areas of work previously not considered or not recognised as library cataloguing.

Chapter 6. Librarianship on the Leading Edge: The Outward Trajectory of the Cataloguing Profession

Introduction

In contrast to Chapter 5 which discussed the expansionary trajectory of Blended Cataloguing as responses to changes within the library and from the perspective of those currently in the profession, this chapter focuses on the outward trajectory of the cataloguing profession.

A thread running through the prior publications reveals that some elements of cataloguing work post-digital revolution, when considered in a broader information space, would be beyond the traditional scope of the cataloguing profession. The appearance of these non-traditional elements might be what caused the “chaotic landscape” of the cataloguing profession (discussed in Chapter 5). This chapter attempts to clarify and refocus the direction of library cataloguing by synthesising pertinent observations from the prior publications. Three broad areas are identified and are called “emergent areas.”

After a discussion of the origins and the scope of these emergent areas, a discussion follows to place them in the context of the “Participatory Library,” a framework of library services first introduced by Lankes, Silverstein and Nicholson (2006) and subsequently appended to the Buckland framework by Bonfield (2014).

While these practices have become more commonplace in the cataloguing profession in recent years, the areas of work lack labelling within the library community and, therefore, are not widely recognised. Contrasting the role of the Blended Cataloguer in the context of the Participatory Library highlights the significance of incorporating these emergent areas, already performed by library cataloguers as well as by workers who ally themselves in other professional fields, into the professional identity.

The Outward Trajectory of the Blended Cataloguer

The three emergent areas can be summarised as data curation, knowledge creation, and policymaking. These three areas are discussed below.

Emergent Area 1: cataloguer as data curator

Cataloguing has its origin in the textual information presented on book pages. As libraries began collecting a broader variety of materials, non-textual materials were included. In the Blended Library, the scope was further expanded to data embedded in digital library resources.

The work of devising new methodologies to handle cataloguing of an expanded set of materials, such as decomposition and reassembly, is itself quite significant. But, these new methodologies enable much more than producing the library catalogue. Cataloguers are now also concerned with constructing new ways of compiling, recording, collocating, and presenting data previously unavailable in the library catalogue. This curation activity involves beyond transcribing and recording data at their face value, largely as instructed in traditional cataloguing rules.

In traditional cataloguing, cataloguers create library metadata, for the most part, from examining the "item in hand," that is, directly transcribing and recording data from the bibliographic item. In the modern library catalogue, in addition to keeping track of each physical item, the bibliographic data serve to describe both the bibliographic item and the underlying intellectual work and the expression pertaining to the item. However, the source of information about bibliographic works and items almost always exceeds the item itself. The electronic materials that libraries began to collect provided further impetus for cataloguers to look beyond the book, where data curation came into play.

Among the areas of bibliographic description, AACR2 rules stipulated specific physical locations of the item to be the "chief source" for the title, edition, publication, and series area. For other areas of description, such as the notes area and terms of availability, information could be taken from any source. RDA relaxed the specifications when the terminology changed to "preferred source" and offered a longer list of alternatives. Also, instead of specifying a different set of sources for every material format, RDA simplified the process by grouping library materials into three groups: items of print origin, moving images, and then a third catch-all group of "all other sources." For these "all other sources" (RDA 2.2.2.4) the sources of information can be very broad — label, embedded metadata, or any "source where the data is formally presented." This presented the opening for the work described in the prior publications. Together, these publications lay the intellectual and practical groundwork for recognising the work of data curation.

[Paper 2] analysed the state of cataloguing at a critical juncture of the digital revolution when cataloguing rules, the library bibliographic model, and linked data technology converged. It concluded by anticipating a future role for the cataloguer as “data curator.” The catalogue of historical photographs discussed in [Paper 1] showcases this curator role. In addition to linking metadata of digitised images to existing catalogue records of the physical items, the descriptive metadata in the existing records were also revisited. One of the elements of enrichment was geographic metadata. Existing geographical subject headings pointed variously to the level of town, village, or neighbourhood. More precise geographic location at the level of building, landmark, or modern street address would be added. After setting up a data structure for street address, finer details were added to each photograph based on existing narrative descriptions, historical knowledge, and historic maps. These addresses were then geocoded in bulk through an automated process through which geographic coordinates were obtained. Also by automation, the geographic coordinates were written into the embedded EXIF metadata of the digital images, as well as reassembled, along with the detailed addresses, into the VRA Core metadata accompanying the digital images. This data curation activity ultimately enabled a novel cartographic presentation of the collection. [Paper 1] described the automated process that created the cartographic catalogue, an early experiment performed on overlaying item descriptions in the Google Maps interface. The ability to display item descriptions directly on a map with an image thumbnail filled a frequent request from library users and, as a library service, was a significant innovation, four years ahead of Google itself rolling out the historical imagery capability.

As extensions of this project, there are many possibilities for data curation. For example, the same collection included historical photographs of the rail network in the region. Researchers would often need to visit various parts of the library, or multiple libraries, for periodicals, books, maps, and patent information. It would be possible for cataloguers to acquire information about locomotives, train stations, rail lines, equipment, architectural structures and buildings, incidents depicted, etc., and present them together on a single interface. However, to offer these data as a library service, librarians would need to assess what was useful, what filled an existing demand, what filled an anticipated demand that library users might not yet be asking for, and what would drive innovation in library service. In traditional reference, public services librarians draw the balance: they do not overwhelm library users with extraneous information, but they also reasonably anticipate the user's next question. Drawing boundaries around the curation of data has been the art of reference service. Anticipating the need to search and filter subsets of metadata in different ways, cataloguers can now create the infrastructure for a rich set of metadata in a ready state to be deployed flexibly and dynamically to meet a large range of library service needs.

Emergent Area 2: cataloguer as knowledge creator

A natural entryway for cataloguers to enable knowledge creation is modelling metadata in new ways that clarify and refine the existing bibliographic data structure, as well as enabling bibliographic metadata to interoperate with non-library data. Cataloguers would become tool creators for their own curated data. These interplays enable novel representations, visualisations, and data processing methodologies which propel the creation of new knowledge.

Traditional cataloguing structured bibliographic data for the library catalogue. Although the functions of the library catalogue evolved over time, the primary purpose of bibliographic descriptions has been to satisfy the primary user tasks under FRBR: find, identify, select, and obtain. These user tasks limited the scope of cataloguing to the library's bibliographic universe. However, as cataloguers are capable of curating non-library data for both library and non-library use, the potential for a two-way interplay already exists.

A significant step took place in 2016 when LRM, the library reference model that succeeded FRBR, included one additional user task: explore. "Explore" is what gives cataloguers the licence to practise bidirectional data contextualisation. In previous reference models, contextualisation was limited to terms related to subject headings and name authorities. The LRM user task "explore" expanded the scope to include any library resource: "to discover resources using the relationships between them and thus place the resources in a context" (Riva, Le Bœuf and Žumer, 2017, p. 8) How can cataloguers help with, enable, and enhance this new user task? "Explore" is broadly applicable to contextualisation of bibliographical data in any relevant context. The curated data in [Paper 1] recontextualised the catalogue of historical photographs from a text-based list to a cartographic presentation. This new display offer library users a new dimension of service. New knowledge about the geographic distribution and relative distance and proximity was also created visually.

A more complex situation was presented in the reference enquiry in [Paper 3]. The initial enquiry was far from a known item search. The users, seeking a specific song by Franz Schubert, had several contextual clues surrounding the performer in a sound recording, and a vague idea about a single moment of musical content in the song. The users came up empty after consulting a list of Schubert's compositions, the title index in a volume of Schubert's collected songs, a list of Schubert songs by genre, and a list of Schubert songs by instrumentation and voice type. The analysis in [Paper 3] showed how these bibliographic tools as well as sophisticated searches in library catalogues all failed, but how this Schubert song was eventually identified using a combination of bibliographic and non-library metadata. In hindsight, these bibliographic tools were centred on the work, but not the performer. Bibliographic data would be helpful for locating the score once the song was identified. But what helped identify the song in the first place involved a trained librarian deliberately drawing on data from outside the library, in this example, from Discog, a specialised database for sound recordings where performers were given more emphasis.

[Paper 3] also posited that computers could be used as an analytical tool, a tool that would access a broader array of data sources for user exploration. In the paper, the users presented a musical clue involving a melodic high point in the song (which was off by a half scale step), and two "look-alike"/"sound-alike" clues: the performer "Strada" or "Estrada" (for which Google correctly suggested "Stader"), and a suffix of "-lein" (which was actually "-keit"). Addressing these contexts is now possible through error detection and lexical disambiguation techniques being developed today.

Another possibility given in [Paper 3] is contextualising medium of performance, which is traditionally recorded in bibliographic data, but requires remodelling the area of description to make the data machine-actionable. Medium of performance was chosen for demonstration in [Paper 3] and [Paper 6], with some urgency, because it is "typically the point of departure" in "the way musicians think" (Ostrove, 2001, p. 102), and "is a frequent starting point in their search for music resources, and an identifying element of musical works and expressions" ([Paper 6], p. 1). Regarding the theoretical aspect, [Paper 3] analysed the classes of issues in bibliographic data that would necessitate a particular area of description to be remodelled. Meanwhile, [Paper 6] presented a detailed reworking and proposed a new model for medium of performance description. Putting the theory into practice, [Paper 6] detailed specific issues with the dispersal of medium of performance data in the bibliographic data structure, and pointed out that current developments in bibliographic data modelling continued to align with the RDA/MARC structure, akin to a reformatting rather than a remodelling, which would not enhance machine-actionability. The new data modelling of the medium of performance area, when contextualised with the musical works and expressions, will be capable of providing new insights into a broad array of topics, such as the evolution and transmission of musical instruments, the nature of instrumentation in the creative process, the consideration of instrumentation in concert programming, the roles particular instruments play in a regional culture, the roles of individual players and singers in the performance context. It may also facilitate other new digital humanistic studies such as trends in conservatory education (Sasser, 2021), bringing new empiricism into musicological studies (Huron, 1999), and using usage analytics for examining library usage and user behaviour. More specifically, [Paper 6] suggested layering the metadata from this new medium of performance model on to a suite of optical music information retrieval and musical content analysis stack (Lewis *et al*., 2022). As anticipated in [Paper 4], the end result of constructing suitable data models that integrate the primarily text-based bibliographic data with multimedia metadata will spur novel modes of creative expressions and knowledge creation.

With subject matter knowledge, cataloguers are in a great position to engage in work beyond producing the library catalogue, as far afield as machine learning, natural language processing and music information retrieval, and their applications in digital humanities, as well as in areas closer to home such as remodelling bibliographic data.

Emergent Area 3: cataloguer as policymaker

Library cataloguers have been active in policy discussions surrounding cataloguing standards and bibliographic data models. However, they can become much more involved in the broader information space.

[Paper 2] analysed the convergence of technology, cataloguing rules, and the willingness of the library community to embrace linked data technology to make library data available on the open web. But how has the library community been involved with shaping linked data technology and its implementation? While developing BIBFRAME and its extensions moved toward establishing a gateway for exposing library data, the library community has made few efforts to become participants in the ecosystem. How are librarians, especially cataloguers, part of this larger community, when, as stated in [Paper 2], “resource description of the future will be done on the Web, for the Web”? Most progress so far has gone toward the first part. While laying the groundwork to expose library data has its own benefits such as increased efficiency through reuse and enrichment, the library community has not fully considered the linked data ecosystem itself as a consumer of library data. In the past decade, the availability of bibliographic and authority data has been greatly increased. But who will be the users? Have libraries take part in creating a superstructure that makes use of library data? The role of “data curator” discussed above takes data into the library from the ecosystem, whereas the role of “policymaker” looks outward as library data are pushed out into the ecosystem.

[Paper 3] makes a clear case about the lack of connecting library data structure to user behaviours in the open web. The open web is a much more messy place, and users of the open web are often messy searchers. But given complete and well structured data, technology is there to help steer users. With messy searchers, [Paper 3] anticipates “linked open data could enable machines to overcome the uncertainties” such as errors in names, prefixes and suffixes, nationality or history of people and objects, relative track sequences in a sound recording, and qualitative understanding of musical contours and mood. With library cataloguing still entrenched in the RDA/MARC standards, who will redirect broader, more general data policies to enable these functions?

The data model implementation shown in [Paper 6] is one such instance. In the modelling of the medium of performance area of description, a number of the listed refinements are out of the ordinary for a traditional library catalogue, but are of immense importance to users. Traditional cataloguing recorded a limited amount of information—the instruments, voices and number—primarily because these pieces of information are readily available on the chief source of information of music publications. As the library catalogue was not originally built for the digital environment, there are considerable conflicts in its fundamental data modelling, such as ambiguity and atomicity issues, which are discussed in [Paper 3] and in [Paper 6].

According to these two papers, it is important to curate data for these refinements for medium of performance, and just as important is for cataloguers to build library service data models in a way that meets user demands, at the same time being compatible and extensible for broader use in the linked data ecosystem as indicated in the “Next Steps” in [Paper 6]. Creating library data but failing to develop the data model would not be a wise strategic path going forward. And, there, cataloguers should emphasise the work on data modelling as a policy issue, rather than a technological issue.

In fact, there are still opportunities to participate in policymaking in nascent areas. An example is electronic music scores. [Paper 5] offered foundational policy work on the direction of collecting and handling of this relatively new and popular material format. While cataloguers have been struggling to devise ways to catalogue and provide access to electronic music scores (Peters, 2019), the conversation, as advocated in [Paper 6], should be about the entire life cycle of this material. Libraries will not purport to be experts in musical contents, but could be influential in developing the infrastructure for the life cycle of electronic music scores. [Paper 6] noted that, while librarians, especially cataloguers, have taken keen interest in the development of music encoding standards, the format of the electronic music scores and the platform on which they are created and consumed have not been standardised. It is an opportunity and an important area to become involved in, and librarians can steer it in an open and sustainable direction. A lesson might be learned from the development of electronic text, where the library community’s early embrace of TEI fell by the wayside and the format gave way to the rise of proprietary DRM (digital rights management) in electronic books which deprived the format’s portability to this day, with enormous consequences. It would be unwise for the library community to repeat the mistake of abandoning participation in developing a standard for a new material format. [Paper 5] attempts to lead the way by establishing a set of functional requirements for the file format for electronic music scores. From the policy perspective, cataloguers could be the driver behind not only the cataloguing of electronic scores, but also behind the surrounding infrastructure: the material format, as well as the creative, publication and archival platforms. An unlikely application appeared in Hawkins (2021) in which the author cited the metadata policy work discussed in [Paper 6] as essential to the livelihood of independent professional musicians.

Finally, to facilitate knowledge creation and data curation, more emphasis needs to be paid to the policy direction of cataloguing from populating the bibliographic catalogue to capturing the totality of the life cycle. With such a goal, cataloguers will need not only create that data, but also build the data models to house them. The resulting outward looking cataloguing profession would capture the spirit of [Paper 1], too early to be effable at the time, of “digitizing everything.”

The Blended Cataloguer in Post-Buckland Frameworks

The evolutionary stages of library services are succinctly summarised by Nancy Nyland in her review of *Blueprints for 2025*: “Libraries are seeking new language that more accurately reflects their evolution from the warehouse model of the paper library through the stages of the electronic library, still housed in a physical space, the online library housed on the Web, and, in the future, the library of connection and creation, or the participatory library” (2014, p. 225).

In his *Manifesto* (p. 4), Buckland emphasised the significance of distinguishing between the means and the ends, and argued that the purposes of and justification of library services should not be confused with the techniques and technologies adopted for providing them. He astutely pointed out that “the long period of relative stability from the late nineteenth century to the 1970s in the means for providing library service” made it easy to blur the distinction between means and ends. He went on, “Alternative means do need to be explored aggressively otherwise the options will not be known” (p. 4), implying the outward trajectory of library cataloguing discussed in this chapter—exploring additional means through expanding the repertoire of cataloguing. Only through this exploration can libraries connect to the ends. They must examine their newfound capabilities, and allow these newfound capabilities to improve and drive evolution of library service.

Fifteen years after Buckland’s *Manifesto*, Lankes, Silverstein and Nicholson (2006) introduced the “participatory library” framework. They noticed that Web 2.0 shifted user behaviours to favouring conversational forms of communication in social media, and web search engines became popular as the primary information seeking interface. In their groundbreaking work, they placed libraries and library services in a “participatory framework,” which they called Library 2.0, where the focus of library services was a space that enabled knowledge creation and dissemination through community interaction. Libraries became part of a “participatory network.” A definition of “participatory librarianship” emerged as being “facilitators and actors in conversations” in the broader information technology landscape.

Rather than being a collector of information and a passive answerer of reference questions, the library becomes a place of community participation, and the participatory librarian is a leader, a model participant who drives the participatory network. In the “test bed” environment in Lankes, Silverstein and Nicholson (2007), libraries would provide public participatory host platforms such as “blogs, wikis, discussion boards, RSS aggregators, and the like.” And since libraries would be running participatory platforms, librarians would have to have the technological know-how to “provide the open source software and consulting support to implement features locally.” The involvement of libraries in developing the vertically integrated infrastructure of electronic music scores proposed in [Paper 5] is such an example.

The missing piece from Lankes, Silverstein and Nicholson (2007) is the “how.” How are librarians going to get there? What are these new skills that the participatory librarian would possess? How would librarians acquire them to support these new technical operations? Cataloguers are already deeply invested in library technical operations and are natural candidates to take on these roles. While the studies in the prior publications were done without an underlying goal of espousing a participatory framework, the three emergent areas fulfil key technological requirements undergirding a participatory library, as stated in the “roadmap” in Lankes, Silverstein and Nicholson (2006): hosted community services, knowledge base, digital repository, federated search, and recommender system.

Data curation in the participatory library

Reaching the participatory framework is a great technological achievement. But how can libraries maintain and sustain it and participate successfully and meaningfully in this environment? Libraries will benefit from a broader goal of being not only actors in this environment, but also developers and drivers. 20 years since Buckland’s *Manifesto* and six years since the term “participatory librarian” was coined, another set of authors came together to examine the direction of libraries. In *Planning Our Future Libraries: Blueprint for 2025*, Brett Bonfield (2014) called Buckland’s stages of libraries the groundwork that enabled the ultimate “Participatory Library.” And as shown in this chapter, data curation will play a significant role in the input into the participatory system: what areas of data will be linked to the traditional library catalogue to make the participatory system prosper? Lankes, Silverstein and Nicholson (2007) gave a hypothetical scenario where the library catalogue was transformed into a platform where services would be presented with audio and video as well as text, library users would answer each other's catalogue queries and leave annotations to library items, and circulation data would be factored into search results relevance rankings. What they are describing is integrating the library catalogue into a data curation platform. Embedded metadata described in [Paper 4] would be an integral part of the multimedia library catalogue display; the user community would help identify the photographs and input the geographic information in [Paper 1]; the data model of [Paper 6] would readily be converted into a crowdsourcing mechanism where library users would help provide the medium of performance information. In these cases, the library serves as a site that enables user participation, at scale, in verifying or disputing the inputs.

Knowledge creation in the participatory library

One of the goals of the participatory librarian is to be a facilitator of knowledge creation. The creation and maintenance of innovative platforms is key to this part of the work. To host a wiki or a blog is more than hosting the blank page and then waiting for the contents to appear. The host would also need to construct a data structure where curated data could be employed on these platforms. Skill sets such as data modelling, crosswalking, and data visualisation would be key to achieving this goal. In fact, the Chicago Crime map, Figure 2 in Lankes, Silverstein and Nicholson (2007), is a rudimentary form of what [Paper 1] demonstrated: while the Chicago Crime map shows pegs where incidents occurred, the pegs shown in [Paper 1] pop up a display of bibliographic data, image thumbnails, and links for accessing the full metadata and high resolution images. That cartographic platform would be one that is worth pursuing, and has been put into practice, thus far, mainly by commercial players. [Paper 5] further sets up a future possibility where the library is a leader-participant in the full life cycle of the electronic music score, by supporting a creative, publishing, and archival platform. Through these platforms, users themselves could create a community of mutual inspiration for further artistic creations and research.

Policymaking in the participatory library

However, a study of participation in libraries by Nguyen, Partridge and Edwards (2012), revealed that policies are often the chief impediment towards making these transformations of library services. This highlights the importance of policymaking. Achieving a Participatory Library involves going two steps forward: creating an entirely new infrastructure to support a platform and data format for future library data that have yet to be created. These technology areas are traditionally outside of the perceived purview of the library. Taking such a large leap involves libraries, especially cataloguers, to participate at the policy level, both inside and outside the library. Hand in hand with this policy participation is the re-imagining the professional identity of the library cataloguer, expanding to encompass titles such as the “Web Services Librarian” and “Metadata and Data Curation Librarian” illustrated in Martin and Sheehan (2018). While cataloguers, who are already steeped in technical operations, move into these roles more readily, it is incumbent on the cataloguing profession to define the scope of the profession and the professional identity, and make aware to those who are already working in that field that they may identify themselves as such.

The Cataloguing Profession

It has long been the prediction that future libraries will dematerialise, and the nature of libraries will expand beyond the “book museum” (Rogers, 2013), from one of “hardware” to one of “software” (Rundle, 2012). As early as in 1992, Buckland’s *Manifesto* described the 21st-century Electronic Library where the dematerialized library provides automated self-service of “open stacks” and “public catalogs,” with the emphasis shifting the purpose of libraries from collections to services. Supplementing this prediction, in his forward-looking 1999 work *Library Services in Theory and Context*, Buckland could foresee the change in librarianship based simply on how many schools and programs were changing their name from “library science” to “library and information science,” a broadening of scope from the concrete sense (paper and cardboard) to the abstract sense (information-bearing objects). He could see the centrality of information retrieval, and that technical principles would remain relevant (including indexing, traditional cataloguing, classification, content analysis and description, techniques of storage, and techniques of retrieval), but also foresaw that the context would change when applying these principles to the broader management of knowledge and representations of knowledge, of materials outside bibliographic control. However, as of 1999, although Buckland could foresee impending upheavals due to imminent changes in technology, he could not foresee exactly how library services would change, or how the change would manifest in library cataloguing.

By 2014, several authors in *Blueprints for 2025* offered more concrete predictions. Bonfield (2014) anticipated the outward expansion into the Web 2.0 “participatory network” appended to Buckland’s framework; Rundle (2014) anticipated an expansion outward into “free-range librarianship” where public service librarians became independent of information service platforms, devices, and even physical locations; Hodge (2014) anticipated libraries taking back ceded grounds from commercial vendors and, revising Buckland’s central purpose of the library, becoming a site where data and information would be transformed into practical knowledge for library users.

As of the writing of this critical commentary in 2023, many of these predictions remain aspirations, as libraries continue to be suspended in the state of the Blended Library. However, for the cataloguing profession, the groundwork has been laid for the next stage, as evident in the expansionary and outward trajectory demonstrated in the prior publications. The three emergent areas represent a departure from the expectation of the cataloguer in a hybrid setting in which the mixture of paper and electronic library materials are catalogued in with the traditional standard-based approach.

The material scope, library services, skill sets and expertise commanded by the Blended Cataloguer have the potential to make all the above a reality. The question is one of identity and politics.

In the *Manifesto* (1992), Buckland stated, “The central purpose of libraries is to provide a service: access to information.” Earlier, he wrote, in *Library Services in Theory and Context* (1988), “The techniques and knowledge associated with librarianship appear to be more or less applicable to a range of activities outside of library service as well. In other words, the techniques and knowledge associated with a range of cognate activities appears to coincide or at least overlap to an important degree” (p. 24). This chapter has shown otherwise. Rather than viewing certain cataloguing activities as outside of library service, cataloguing is well positioned to subsume these cognate activities into its own professional identity, in a virtuous cycle of innovation as advocated by Hodge (2014). This expanded identity can drive policies in libraries and in the information space toward the participatory framework envisioned in Bonfield (2014). In this participatory framework, Buckland's central purpose of libraries will be revised, providing not only access to information, but also to information tools. In other words, Buckland's concern with the conflation of the means and the ends of library service will become moot. The new purpose of libraries will provide to users the means and the ends; thus expanded, cataloguing *is* library service. In 2012, Hugh Rundle wrote in his blog, "Libraries are a technology for large-scale inter-generational transfer of knowledge and culture…We can envisage the library as a platform for enabling innovation, learning and cultural development to occur in our communities without the need for capital." This critical commentary argues that the capital is in fact the information infrastructure constructed by cataloguers, the platform that supports the free-range librarian as Rundle later modelled his 2014 essay. Cataloguers, in their Blended identity or whatever appellation may come next, will be the central players in developing and sustaining this infrastructure.

Chapter 7. Conclusion

Chapter Summaries

Chapter 2 summary

Chapter 2 drew on two common threads among the prior publications. These two threads situated the new term “Blended Library” within the library services framework of Buckland (1992) to describe the state of persistent coexistence between paper and electronic library materials and services, and presented the argument for the urgency to expand the professional identity of library cataloguers to the “Blended Cataloguer” through embracing the generalised Decomposition-Assembly Approach to cataloguing and metadata work. Building toward this expanded professional identity, Chapter 2 laid out the four areas of intersection between the Blended Library and library cataloguing: the process, the product, the person, and the profession, which would be the subject of the next four chapters.

Chapter 3 summary

Chapter 3 examined interactions between cataloguing practices and the nature of library materials. After a technochronological exposition of cataloguing practices, discussion focused on the technical and policy limitations of traditional approaches to adapting cataloguing to electronic materials. Drawing on specific adaptation efforts in the prior publications, an observation was made pointing to a departure from traditional standard-based practices to the emergence of a more generalised approach to cataloguing, the Decomposition-Assembly Approach, which would be discussed in detail in the next chapter.

Chapter 4 summary

Chapter 4 described the state of the library catalogue as it becomes subsumed in the broader technological landscape, and, consequently, the shift of library cataloguing toward supplying metadata to the broader information ecosystem. The Decomposition-Assembly Approach was described in fuller detail in this context as a practical approach for the “Theory of Metadata Enriching and Filtering” exposited by Alemu and Stevens (2015), accompanied by use cases drawn from the prior publications. These use cases made evident that this approach is not only suitable for producing the library catalogue, but also enables novel applications for library services. To enable support for these applications, the cataloguer would be required to curate the data and develop the data structure, which represents a shift toward a prospective mindset that drives the cataloguer toward becoming a knowledge-producing participant in the information ecosystem. The expansion of the scope of the cataloguing profession into new emergent areas would be further developed in Chapters 5 and 6.

Chapter 5 summary

Chapter 5 placed the trajectory of the cataloguing profession in the prospective context of the Blended Library developed in the previous two chapters. The expansion of cataloguing practices and the functions of the library data was placed into the context of the cataloguing profession as a whole in terms of standards, knowledge, and expertise, characteristics that define a profession. This profession of data-oriented “Blended Cataloguer” was distinguished from the public service-oriented “Blended Librarian” coined by Bell and Shank (2004). Drawing on use cases in the prior publications, the expanded scope of the Blended Cataloguer was shown to be capable of reclaiming librarianship from vendors, publishers, and internet search engines, the three existential anxieties identified by Bell and Shank.

Chapter 6 summary

Chapter 6 mapped the expanded practices of the Blended Cataloguer, as demonstrated in the prior publications, to three emergent areas: data curator, knowledge creator, and policymaker. As a profession involved in these areas, the Blended Cataloguer was shown to fill the practical aspect omitted in the “Participatory Library,” a future extension of Buckland’s framework proposed by Lankes, Silverstein and Nicholson (2006). The Blended Cataloguer as developer and sustainer of library technology infrastructure was then in this participatory space in relation to the organisation of libraries proposed by Hodge (2014) and alongside the “Free-Range Librarian” envisioned by Rundle (2014).

Research Questions

The prior publications and this accompanying critical commentary submitted for this Ph.D. by Prior Publication traced the expanding scope of the library cataloguer in a time of rapid technological change in the second decade of the twenty-first century. Through elucidating the themes and threads the research questions have been addressed as follows:

In what ways has the cataloguer's role transformed and in what ways will it continue to transform as the library community adopts linked data technologies and shares library data in the open information space?

The cataloguer's work has been stable for several decades and a desire to maintain the status quo only grew stronger. However, rapid technological change brought new digital materials and new electronic services to libraries. In this Blended Library, cataloguers have attempted to adapt to the new materials, but, to meet new service demands, the cataloguer's work must necessarily move beyond producing the traditional library catalogue. The prior publications have shown that cataloguing work has already entered new areas, and this critical commentary highlights these new roles that cataloguers have taken on—data curator, knowledge producer, and information policymaker. Together, these roles form a re-imagined professional identity of the Blended Cataloguer. The Blended Cataloguer will provide the information infrastructure support necessary for library data to be shared in the open information space. These roles will also allow cataloguers to step into leadership roles that shape the information space itself, propelling libraries to the forefront of the future information ecosystem.

In what ways and to what extent will this transformation necessitate revisions and innovation of the cataloguer's knowledge, skills, and tools?

Technological change also brought new available tools. To show practically how to optimise the use of the new available tools, this critical commentary illustrated the Decomposition-Assembly Approach as a common thread throughout the new roles of the Blended Cataloguer. By practising this approach, the Blended Cataloguer adds a new set of skills that include aspects of data science applicable to library operations, such as computer programming, data processing methods, and data analysis and statistics. Just as Buckland (1988) noted the addition of "information" to "library science" (p.16) the Blended Cataloguer could be said to be a practitioner of library, information, and data science. Beyond technical skills, the Blended Cataloguer identity will also require management and leadership acumen. As the new roles involve leadership in the information ecosystem, the scope of work will encompass a broad spectrum of stakeholders beyond the immediate library user interaction or the immediate institution; the location of work will emerge from the back office to a broadly staff- and public-facing environment. While welcomed in the traditional setting, these capabilities will become highly desirable.

In light of these transformations, can a practical model of cataloguing be developed?

The Decomposition-Assembly Approach is a suitable practical model in light of the transformations of professional identity. This critical commentary has shown that the limitations of the century-old standard-based approach to library cataloguing practice as technologies and the information ecosystem evolved into a blended space in recent decades. Decomposition-Assembly represents a new, professional identity-defining approach for the Blended Cataloguer that is capable of expanding the scope and services provided by library cataloguing, exemplified by the re-contextualized map-based catalogue at the Queens Library, while not abandoning, but fully utilising, the high-quality metadata of traditional cataloguing and the familiar product that is the library catalogue, exemplified by the compatibility of OnVIE with MARC.

Strengths, Limitations, and Future Directions

While the influence of the prior publications has been presented in Chapter 1, this section discusses the prospects of materials synthesised in this critical commentary. The two threads developed in this critical commentary engaged in dialogue with a long-running area of scholarship on visions of the future library alongside technological change, and offered a practical approach to realising these visions. The concept of the Blended Library was inserted into the stages of evolution of library services to describe an expected perpetual state which was, three decades ago, anticipated as a transitional period, and, therefore, omitted in Buckland's framework in his 1992 *Manifesto*; the Decomposition-Assembly Approach provided a concrete, practical approach to Alemu and Stevens's theoretical model for digital library metadata. This approach was also shown to be central to enabling the new roles—data curator, knowledge creator, policymaker—that form the core identity of the profession. This new professional identity would complement the more recent scholarship of the future library by addressing foundational information infrastructure support for the library as an institution, the library's organisational structure, and the mode of library service envisioned in the 2014 essays by Bonfield, Hodge, and Rundle.

The transformation of professional identity from the traditional cataloguer to the Blended Cataloguer is not a trivial one, requiring radical reconsideration on the conceptual, technical, and organisational level. The most notable strength of the Decomposition-Assembly Approach, its flexibility, defies the one-size-fits-all model of standardisation familiar to, and, perhaps, enjoyed for over half a century by many traditional cataloguers. Taking on new roles of data curator, knowledge creator, and policymaker requires developing expertise and confidence at the individual level for day-to-day practice, and leadership in the broader information ecosystem. However, the history of inertia in traditional cataloguing practices suggests challenges ahead, and other players in the information sector will continue to present competition.

Glimpses of these challenges were already seen in the prior publications, yet issues presented in these prior publications continue to be cited worldwide in an array of applications in technical areas such as electronic music scores, information retrieval, metadata standards, and discovery interface design, as well as in humanistic pursuits such as digital musicology, social history, and media arts. The common threads brought together and terminologies coined in this critical commentary aim to enable a clearer, unified path not only toward these areas of research, but more broadly spark virtuous cycles of innovation in librarianship.

Concluding remarks

Speaking on a panel of past presidents of the Music Library Association in the United States, Geraldine Ostrove, a leading authority on subject and genre/form who retired as the Senior Cataloguing Policy Analyst in the Policy and Standards Division of the Library of Congress, spoke on the future of librarianship, recalling a situation in 2007: "Even then, the name we give ourselves in our profession was becoming obsolete...Things change. And maybe that our jobs as we see them now change dramatically" (Music Library Association, 2016, minute 54-55). As the future unfolded, library cataloguers, digital musicologists, computer engineers, and data scientists all became part of the same conversation. In an interview, University of Chicago economist Steven Leavitt (2022) spoke of his preferred “approach” to big data research as “the possibilities that come when you use rich data in a way that nobody had ever imagined it would be used.” For Leavitt, this approach also means using “data that were designed for one purpose…in a completely different way to try to answer questions we’ve always cared about.” Will cataloguers rise to the challenge?

Reference List

Alajmi, B., and ur Rehman, S. (2016) ‘Knowledge organization trends in library and information education: assessment and analysis’, *Education for Information*, 32(4), pp. 411-420. Available at: https://doi.org/10.3233/EFI-160084

Alemu, G. and Stevens, B. (2015). *An emergent theory of digital library metadata: enrich then filter*. London: Chandos Publishing.

Anglo-American cataloguing rules. 2nd edn, rev. 1988. (1988) Chicago: American Library Association.

Bair, S. (2005) ‘Toward a code of ethics for cataloging’, Technical Services Quarterly, 23(1), pp. 13-26. Available at: https://doi.org/10.1300/J124v23n01_02

Bawden, D., Robinson, L, Anderson, T., Bates, J., Rutkauskiene, U. and Vilar, P. (2007) ‘Towards curriculum 2.0: library/information education for a web 2.0 world’, *Library and information Research*, 31(99), pp. 14-25. Available at: https://doi.org/10.29173/lirg49

Bell, S. and Shank, J. (2004) ‘The blended librarian: A blueprint for redefining the teaching and learning role of academic librarians’. *College & Research Libraries News*, 65(7), pp. 372-375. Available at: https://doi.org/10.5860/crln.65.7.7297

Bonfield, B. (2014) ‘Redesigning library services again: revisiting Buckland’s Manifesto’, in K. Leeder and E. Frierson (eds) *Planning our future libraries: blueprints for 2025*. Chicago: American Library Association, pp. 3-14.

Brugger, J.M. (1996) ‘Cataloging for digital libraries’. *Cataloging & Classification Quarterly*, 22(3-4), pp. 59-73. Available at: https://doi.org/10.1300/J104v22n03_05

Buckland, M.K. (1988) ‘Bibliography, library records, and the redefinition of the library catalog’. *Library Resources and Technical Services*, 32(4), pp. 299-311. Available at: https://www.learntechlib.org/p/141330/

Buckland, M.K. (1992) *Redesigning library services: a manifesto*. Chicago: American Library Association. Available at: https://digitalassets.lib.berkeley.edu/sunsite/Redesigning%20Library%20Services_%20A%20Manifesto%20%28HTML%29.pdf

Buckland, M.K. (1988) *Library services in theory and context*. 2nd ed. Oxford, Pergamon.

Caplan, P. and Guenther, R. (1996) ‘Metadata for Internet resources: The Dublin Core metadata elements set and its mapping to USMARC’. *Cataloging &*

Classification Quarterly, 22(3-4), pp.43-58. Available at: https://doi.org/10.1300/J104v22n03_04

Cerbo, M.A. (2011) ‘Is there a future for library catalogers?’, *Cataloging & Classification Quarterly*, 49(4), pp. 323-327. Available at: https://doi.org/10.1080/01639374.2011.571504

Copeland, A. (2002) ‘Works and digital resources in the catalog: electronic versions of Book of Urizen, The Kelmscott Chaucer and Robinson Crusoe’. *Cataloging & Classification Quarterly*, 33(3-4), pp. 161-180. Available at: https://doi.org/10.1300/J104v33n03_08

Coyle, K. (2000) *Is MARC dead?* Available at: http://www.kcoyle.net/marcdead.html

Coyle, K. (2011) ‘MARC21 as data: a start’. *Code4lib Journal*, 14. Available at: https://journal.code4lib.org/articles/5468

Danskin, A., (2007) ‘“Tomorrow Never Knows”: the end of cataloguing?’ *IFLA journal*, 33(3), pp. 205-209. Available at: https://doi.org/10.1177/0340035207083301

Davis-Brown, B. and Williamson, D. (1996) ‘Cataloging at the Library of Congress in the digital age’. *Cataloging & Classification Quarterly*, 22(3-4), pp. 171-196. Available at: https://doi.org/10.1300/J104v22n03_12

Dillon, M. and Jul, E. (1996) ‘Cataloging Internet resources: the convergence of libraries and Internet resources’. *Cataloging & Classification Quarterly*, 22(3-4), pp. 197-238. Available at: https://doi.org/10.1300/J104v22n03_13

Dobreski, B. (2021) ‘Descriptive cataloging: the history and practice of describing library resources’. *Cataloging & classification Quarterly*, 59(2-3), p. 225-241. Available at: https://doi.org/10.1080/01639374.2020.1864693

Dobreski, B., Zhu, X., Ridenour, L. and Yang, T. (2022) ‘Information organization and information retrieval in the LIS curriculum: an analysis of course syllabi’, *Journal of Education for Library and Information Science*, 63(3), pp. 335-350. Available at: https://doi.org/10.3138/jelis-2021-0057

Gaitanou, P., Andreou, I., Sicilia, M.-A. and Garoufallou, E. (2022) ‘Linked data for libraries: creating a global knowledge space, a systematic literature review’. *Journal of Information Science*, pp. 1-41. Available at: https://doi.org/10.1177/01655515221084645

Geckle, B.J. and Nelson, D.N. (2017) ‘Classifying librarians: cataloger, taxonomist, metadatician?’. *The Serials Librarian*, 72(1-4), pp. 57-64. Available at: https://doi.org/10.1080/0361526X.2017.1320871

Gorman, M. (1978) ‘The Anglo-American Cataloging Rules, Second Edition’. *Library Resources and Technical Services, 22*(3), pp. 209-226.

Gorman, M. (1999) ‘Metadata or cataloguing? A false choice’. *Journal of Internet cataloging*, 2(1), pp. 5-22. Available at: https://doi.org/10.1300/J141v02n01_03

Griscom, R. (2022) ‘Music cataloging and technological change in the 1980s: a personal history’. *Notes*, 79 (1), pp. 21-42. Available at: https://doi.org/10.1353/not.2022.0062

Han, M.J. and Hswe, P. (2011) ‘The evolving role of the metadata librarian’, *Library Resources & Technical Services*, 54(3), pp. 129-141. Available at: https://doi.org/10.5860/lrts.54n3.129

Hawkins, W.M. (2021) *A musician’s livelihood: a qualitative study examining perceived barriers to profitability for independent musicians*. PhD thesis. Northcentral University.

Heilig, M., Rädle, R. and Reiterer, H. (2011). ‘Die Blended Library: Benutzerorientierte Verschmelzung von virtuellen und realen Bibliotheksdiensten’. In B. Bekavac, R. Schneider and W. Schweibenz (eds) *Benutzerorientierte Bibliotheken im Web: Usability-Methoden, Umsetzung und Trends*. Berlin: De Gruyter Saur, pp. 217-242.

Hodge, M. (2014) ‘The constant innovator: a new organizational mode of experimentation’, in K. Leeder and E. Frierson (eds) *Planning our future libraries: blueprints for 2025*. Chicago: American Library Association, pp. 71-82.

Howarth, L.C. (2005) ‘Metadata and bibliographic control: Soul-mates or two solitudes?’, *Cataloging & Classification Quarterly*, 40(3-4), pp. 37-56. Available at: https://doi.org/10.1300/J104v40n03_03

Huron, D. (1999) ‘The new empiricism: systematic musicology in a postmodern age’, in D. Huron (ed) *Music and mind: foundations of cognitive musicology*.

Berkeley: University of California, Berkeley Department of Music. The 1999 Ernest Bloch Lectures: Lecture 3. Available at: https://archive.org/details/huron-1999-ernst-bloch-lectures-music-and-mind-foundations-of-cognitive-musicology/page/n55

Ivey, R.T. (2009) ‘Perceptions of the future of cataloging: is the sky really falling?’, *Cataloging & Classification Quarterly*, 47(5), pp. 464-482. Available at: https://doi.org/10.1080/01639370902832645

Joudrey, D.N. and McGinnis, R. (2014) ‘Graduate education for information organization, cataloging, and metadata’, *Cataloging & Classification Quarterly*, 52(5), pp. 506-550. Available at: https://doi.org/10.1080/01639374.2014.911236

Ke, P., Peng, L., Yuan, Z.Z., Yang, J.X. (2019) ‘Analysis of foreign library science highlights and frontiers, 2018’. *Journal of Library and Data* [柯平, 彭亮, 袁珍珍 and 杨济霞 (2019) ‘2018 年国外图书馆学热点与前沿分析’. 文献], 1(1), pp. 92-107. Available at: https://doi/org/10.31193/SSAP.J.ISSN.2096-6695.2019.01.07

Kochen, M. (1972) ‘Directory Design for Networks of Information and Referral Centers’, *Library Quarterly*, 42, pp. 59-76. Available at: https://www.jstor.org/stable/4306140

Lankes, R. D., Silverstein, J., and Nicholson, S. (2006) *Participatory networks: the library as conversation: executive summary*. Available at: http://hdl.handle.net/11213/9241

Lankes, R. D., Silverstein, J., and Nicholson, S. (2007). ‘Participatory networks: the library as conversation’, *Information Technology and Libraries*, 26(4), pp. 17-33. Available at: https://doi.org/10.6017/ital.v26i4.3267

Leavitt, S. (2022) *People I (mostly) admire: episode 95: The one thing Stephen Dubner hasn’t quit*. Available at: https://freakonomics.com/podcast/the-one-thing-stephen-dubner-hasnt-quit/

Leazer, G.H. (1993) *A conceptual plan for the description and control of bibliographic works*. PhD thesis. Columbia University.

Leeder K. and Frierson E. (eds) (2014) *Planning our future libraries: blueprints for 2025*. Chicago: American Library Association.

Lewis, D., Page, K. and Dreyfus, L. (2021) ‘Narratives and exploration in a musicology app: supporting scholarly argument with the Lohengrin TimeMachine’, *Proceedings of the 8th International Conference on Digital Libraries for Musicology*, pp. 50-58. Available at: https://doi.org/10.1145/3469013.3469020

Lewis, D., Shibata, E., Saccomano, M., Rosendahl, L., Kepper, J., Hankinson, A., Siegert, C. and Page, K. (2022) ‘A model for annotating musical versions and arrangements across multiple documents and media’, *Proceedings of the 9th International Conference on Digital Libraries for Musicology*, pp. 10-18. Available at: https://doi.org/10.1145/3543882.3543891

Library of Congress (2005) *Bibliographic control of web resources: a Library of Congress action plan*. Available at: https://www.loc.gov/catdir/bibcontrol/actionplan.pdf

LinkedMusic (2023) *Recent papers*. Available at: https://linkedmusic.ca/papers/

Mandel, C.A. and Wolven, R. (1996). ‘Intellectual access to digital documents: joining proven principles with new technologies’. *Cataloging & Classification Quarterly*, 22(3-4), pp.25-42. Available at: https://doi.org/10.1300/J104v22n03_03

Martin, E.A. and Sheehan, L.A. (2018) ‘The new “jack of all”: the evolution of the functionality and focus of the academic librarian in new spaces and new roles’, in G.J. Fowler and S.S. Hines (eds) *Challenging the “jacks of all trades but masters of none” librarian syndrome*. Bingley, UK: Emerald Publishing Limited, pp. 67-90. Available at: https://doi.org/10.1108/S0732-067120180000039006

McMillan, G. (1996) ‘Electronic theses and dissertations: merging perspectives’. *Cataloging & Classification Quarterly*, 22(3-4), pp.105-125. Available at: https://doi.org/10.1300/J104v22n03_09

McCallum, S. (2012) *Bibliographic Framework Initiative approach for MARC data as linked data* [Presentation]. The International Group of Ex Libris Users (IGeLU)

Conference, Zurich: 13 September. Available at: https://www.loc.gov/bibframe/pdf/BFI-IGeLU-ppt2003.pdf

Miksa, F. (2009) ‘Chan, Taylor, and the future of cataloging texts’. *The Library Quarterly*, 79(1), pp. 131-143. Available at: https://doi.org/10.1086/593369

Miksa, F. (2012) ‘The legacy of the library catalogue for the present’. *Library Trends*, 61(1), pp. 7-34. Available at: https://doi.org/10.1353/lib.2012.0023

Miksa, S.D. (2021) ‘Cataloging principles and objectives: history and development’. *Cataloging & Classification Quarterly*, 59(2-3), pp. 97-128. Available at: https://doi.org/10.1080/01639374.2021.1883173

Milstead, J. and Feldman, S. (1999) ‘Metadata: cataloging by Any Other Name…’ *ONLINE*, 23, pp.24-31. Available at: https://web.archive.org/web/20000818132720/http://www.onlineinc.com/onlinemag/OL1999/milstead1.html

Morgan, E.L. (1996) ‘Possible solutions for incorporating digital information mediums into traditional library cataloging services’. *Cataloging & Classification Quarterly*, 22(3-4), pp. 143-170. Available at: https://doi.org/10.1300/J104v22n03_11

Moses, B., and Saha, D. (2022) *Data catalogs are dead; long live data discovery*. Available at: https://www.montecarlodata.com/blog-data-catalogs-are-dead-long-live-data-discovery/

Music Library Association (2016) *MLA Presidents Speak: Live in Cincinnati!* [Panel discussion] Available at: https://vimeo.com/170061604

Newcomer, N.L., Belford, R., Kulczak, D., Szeto, K., Matthews, J. and Shaw, M. (2013) ‘Music discovery requirements: a guide to optimizing interfaces’, *Notes*, 69(3), pp. 494-524. Available at: http://doi.org/10.1353/not.2013.0017

Nguyen, L.C., Partridge, H. and Edwards, S.L. (2012) ‘Understanding the participatory library through a grounded theory study’, *Proceedings of the American Society for Information Science and Technology*, 49(1), pp. 1-9. Available at: https://doi.org/10.1002/meet.14504901051

Nyland, N.M. (2014) ‘A Review of “Planning Our Future Libraries: Blueprints for 2025”’, *Journal of Electronic Resources Librarianship*, 26(3), pp. 225-226. Available at: https://doi.org/10.1080/1941126X.2014.939045

Ostrove, G.E. (2001) ‘Music subject cataloging and form/genre implementation at the Library of Congress’, *Cataloging & Classification Quarterly*, 32(2), pp. 91-106. Available at: https://doi.org/10.1300/J104v32n02_08

Page, K.R., Lewis, D. and Weigl, D.M. (2019) ‘MELD: a linked data framework for multimedia access to music digital libraries’, in *Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL),* pp. 434-435. Available at: https://doi.org/10.1109/JCDL.2019.00106

Patrício, H.S., Cordeiro, M.I. and Ramos, P.N. (2021) ‘Ontologias bibliográficas e web semâmtica: limitações e propostas de investigação’, *Páginas a&b: Arquivos E Bibliotecas*, 3(15), pp. 3-29.

Peetz-Ullman, J. (2020) *The difficult task of finding digitized music manuscripts in online library collection*. MLIS thesis. University of Borås. Available at: https://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-23676

Peters, C. (2019) ‘Acquiring PDF Scores for the Music Library: A Progress Report’, *Music Reference Services Quarterly*, 22(3), pp. 131-144. Available at: https://doi.org/10.1080/10588167.2019.1614366

Queens Public Library (2019) ‘Coal chute and power station’, *Digital Archives*, aql:3272, hbf-009970. Available at: http://digitalarchives.queenslibrary.org/vital_node_create/aql_3272

Resource, descriptions & access (2010) Available at: https://original.rdatoolkit.org/

Ruschoff, C.M. (2020) ‘Confronting the digital environment: assessing the challenges to bibliographic control of electronic resources’. *Cataloging & Classification Quarterly*, 58(3-4), pp. 275-284. Available at: https://doi.org/10.1080/01639374.2020.1712510

Riva, P., Le Bœuf, P. and Žumer, M. (eds) (2017) *Transition mappings: user tasks, entities, attributes, and relationships in FRBR, FRAD, and FRSAD mapped to their equivalents in the IFLA Library Reference Model*. Netherlands: IFLA. Available at:

https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/transitionmappings201708.pdf

Rogers, J. (2013) ‘Keynote from NLS6: moving beyond book museums’, *Attempting Elegance*, 24 March. Available at: https://web.archive.org/web/20130420091429/http://www.attemptingelegance.com/?p=1947

Rundle, H. (2012) ‘Libraries as software: dematerialising, platforms and returning to first principles’, *Information Flaneur*, 4 April. Available at: https://www.hughrundle.net/2012/04/04/libraries-as-software-dematerialising-platforms-and-returning-to-first-principles/index.html/

Rundle, H. (2014) ‘Free-Range librarianship: public librarian as park ranger’, in K. Leeder and E. Frierson (eds) *Planning our future libraries: blueprints for 2025*. Chicago: American Library Association, pp. 55-68.

Salaba, A. (2020) ‘Knowledge organization requirements in LIS graduate programs’, *Proceedings of the sixteenth International ISKO Conference*, pp. 384-393. Baden-Baden: Ergon. Available at: https://doi.org/10.5771/9783956507762-384

Sasser, P. P. (2021) *Listening to Metadata* [Conference Presentation]. Described in Phinney S. (2022) ‘Report by Scott Phinney’, *Breve Notes*, 123, pp. 10-14. Available at: http://semla.musiclibraryassoc.org/brevenotes/BN123.pdf

Seaman, D. (1995) ‘Campus publishing in standardized electronic formats: HTML and TEI’, in A. Okerson (ed) *Scholarly Publishing on the Electronic Networks: Filling the Pipeline and Paying the Piper*: *Proceedings of the Fourth ARL Symposium on Scholarly Publishing on the Electronic Networks*. Washington D.C.: Association of Research Libraries, pp. 93-102.

Shera, J.H. (1972) ‘The librarians' retreat from bibliography’, in J. H. Shera *The Foundations of education for librarianship*. New York: Becker & Hayes, pp.185-190.

Tennant, R. (2002) ‘MARC must die’, *Library Journal*, 127(17), pp. 26-27. Available at: https://www.libraryjournal.com/story/marc-must-die

Tennant, R. (2017) '"MARC Must Die" 15 Years On', *Hanging Together: The OCLC Research Blog*, 15 October. Available at: https://hangingtogether.org/marc-must-die-15-years-on/

Thompson, E.H. (ed) (1943) *A.L.A. glossary of library terms, with a selection of terms in related fields*. Chicago: American Library Association. Available at: https://archive.org/details/in.gov.ignca.1252

Turner, R. (2020) 'Analyzing cataloging job descriptions: are cataloging jobs disappearing, changing, or merging?'. *Cataloging & Classification Quarterly*, 58(6), pp. 591-602. Available at: https://doi.org/10.1080/01639374.2020.1795768

Vellucci, S.L. (1997) 'Cataloging across the curriculum: a syndetic structure for teaching cataloging', *Cataloging & Classification Quarterly*, 24(1-2), pp. 35-59. Available at: https://doi.org/10.1300/J104v24n01_04

Wang, Y. and Yang, S.Q. (2018) 'Tecnologies de dades enllaçades (linked data): el que les biblioteques han aconseguit fins ara'. *Item: revista de biblioteconomia i documentació*, 2(65-66), pp. 7-26. Available at: https://raco.cat/index.php/Item/article/view/353381

Wiederhold, R.A. and Reeve, G.F. (2021) 'Authority control today: principles, practices, and trends'. *Cataloging & Classification Quarterly*, 59(2-3), pp. 129-158. Available at: https://doi.org/10.1080/01639374.2021.1881009

Young, H. and Belanger, T. (eds) (1983) *The ALA glossary of library and information science*. Chicago: American Library Association. Available at: https://archive.org/details/alaglossaryoflib00youn

Prior Publications

Paper 1

Digitizing everything?
[Part I] The launch of the digital imaging program at Queens Library
Part II: Piloting metadata creation

Journal of the Library Administration & Management Section
7(2), pp. 5-15, and 8(1), pp. 31-49

Part I: https://scholarsarchive.library.albany.edu/jlams/vol7/iss2/1
Part II: https://scholarsarchive.library.albany.edu/jlams/vol8/iss1/1

JLAMS

A Peer-Reviewed Journal of the New York Library Association

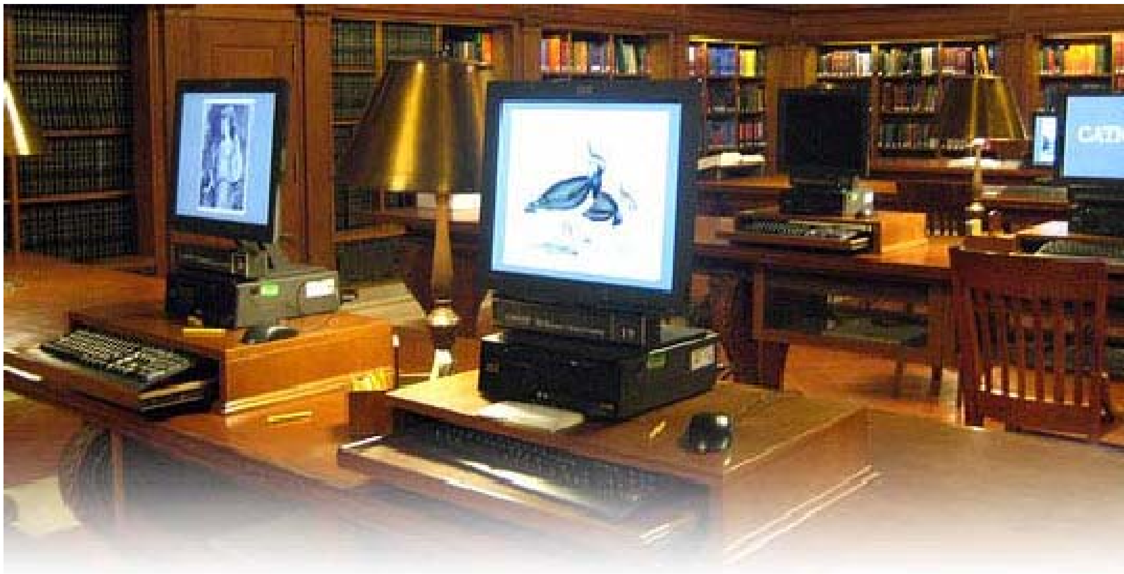
Digitization, Metrics, Dewey & More

Spring
2010-2011
Vol. 7, No. 2

Library Administration and Management Section (LAMS) Officers

To submit an article read the JLAMS editorial policy on our websiteat: *http://www.nyla.org/index.php?page_id=813* Then send an eMail proposal to: naylor@ColonieLibrary.org

New York Library Association
6021 State Farm Road
Guilderland, NY 10284
518-432-6952/800-252-NYLA
email: info@nyla.org
web:www.nyla.org



CONTENTS

DIGITIZING EVERYTHING?

THE LAUNCH OF THE DIGITAL IMAGING PROGRAM AT QUEENS LIBRARY

By Kimmy Szeto

Abstract: In 2006, the Queens Borough Public Library established a Digital Initiative to digitize the contents of its archives. The Initiative was formally launched in 2009, and at that time, I joined the program as its metadata librarian. I served on the committee that was instrumental to the program's metadata policy, workflow design, and implementation. In the course of this work, I frequently revisited the program's development phases and initial launch in order to focus my role. Reexamining this history allowed a deeper understanding of the program's mission. From this perspective, in this paper, I reflect on and examine how the QBPL staff spearheaded the program and launched the administrative unit. I will discuss the key deliberations we undertook regarding the program's institutional impact, major milestones we achieved during the developmental stages, and program-related discoveries we made in the process.

Introduction

This paper addresses both the reservation and the enthusiasm of libraries considering this kind of program. With careful planning, a digital imaging program can be set up on a shoestring budget with a library's existing staff, and possibly with its existing organizational structure. I will show how our library was able to maximize productivity from a small budget and minimal organizational change, and I will also discuss practices that are helping to sustain the program for the long term. QBPL's experience can serve as a paradigm for other institutions that are considering a digital imaging program.

Kimmy Szeto is Assistant Librarian at the Maritime College, State University of New York. He heads cataloging and metadata creation at the Stephen B.. Luce Library and Archives, and manages the Federal Deposit Library Program at the library. His email is kszeto@sunymaritime.edu.

Literature Review

Digital imaging projects are complex and cover a broad range of topics. Kenney and Rieger (2000) serves as a comprehensive practical guide to building and maintaining a digital assets collection. Another comprehensive guide to digitization programs is Hughes (2004), which devotes one half to technical matters and the other half to strategic issues. Greenstein and Thorin (2002), a meta-study of the life cycle of a library's digital program, theorizes on six case studies of large universities.

Major national libraries and organizations involved in digital libraries have set out guidelines for digitization activities. In 2000, the United States Congress established the National Digital Information Infrastructure & Preservation Program for preserving the nation's digital materials. The Library of Congress, charged with carrying out the program, produced a detailed report of the program's history, research, and project planning (LC, 2002). The Council of Library and Information Resources published a strategy paper discussing th e rationales for digitization and institutional impacts of digitization projects (Smith, 2001). On behalf of UNESCO, the International Federation of Library Associations, and the International Council on Archives jointly authored detailed guidelines for digitization projects, including budgeting, human resource planning, and project management (IFLA, 2002). More recently, NDIIPP formed the group called the Federal Agencies Digitization Guidelines Initiative in 2007 to "define common guidelines, methods and practices to digitize historical content." The work of this group has been mostly technical in nature (FADGI, 2010).

The majority of the literature focuses on three areas: the history of digital programs, programs' technical aspects, or digital collections. A few touch on organizational implications: Hunter, Legg, and Oehlerts (2010) describe their collaborative experience when bringing different skills and perspectives from both library and archives worlds; Sennema (2004) recounts his experience launching and implementing a digital media archive on a minimal budget using a new module included in the ILS; Boock (2008) presents survey results for distribution of digitization responsibilities in the organizational structure of ARL libraries.

In contrast to these other works, this paper focuses on the development and organizational support required for the launch of a large-scale digitization program from the middle administrator's perspective. It is my hope that presenting this experience will help other libraries weigh the costs and benefits of launching a digitization program and find the best way to administer it.

Methodology

The founding of QBPL's Digital Initiative and its eventual launch involved many parallel strands. Therefore, instead of using a chronological narrative, I organize significant events, milestones, and major decisions by topic. In each section, I reflect on my experience and discuss the issues that arose in the decision-making process, and analyze how the staff worked through the particular issues. Finally, I offer general, practical suggestion with regard to these issues.

Founding the Digital Initiative

Motivation for the Digital Initiative

The Digital Initiative at the Queens Borough Public Library was born out of a crisis in the mid 2000s. The Library was considering closing or significantly reducing the holdings in its Archives Division (formerly known as the Long Island Division), which collects materials documenting the history of Kings, Queens, Nassau, and Suffolk Counties, the four counties that make up Long Island. Proposals were put forth on ways to dispose of the materials, but one in particular caught the attention of the senior leadership. This proposal was quite radical at the time—it called for digitizing the entire holdings of the Archives Division and making all materials freely accessible on the web.

As conceived, the Digital Initiative aligned well with the mission, vision, values, and strategic directions of the Queens Borough Public Library. At a time when digitization technology and web usage were both taking off quickly, the Digital Initiative would involve technologies that "carry the people of Queens into … the future," provide "rapid and comprehensive access" to the digitized materials, and serve as an online "destination for informational, education, cultural, and recreational needs" (QBPL, 1991). In the remainder of the paper, I treat the further considerations that arose in the continuing development of this Initiative.

Benefits of the Digital Initiative

Affirmation of collection's value.: Investing in digitizing library collections was seen as an affirmation of "continuing value of such resources for learning, teaching, research, scholarship, documentation, and public accountability," and also as an affirmation of the library's stewardship in these areas (Kenney and Rieger 2000, 1). In this sense, QBPL's Digital Initiative brought the Archives to the center of library activities, solidifying its position as an indispensible part of the institution. Once on the verge of being downsized, the Archives would become the leader in digital assets management and online content delivery.

International recognition.: Online materials could be accessed remotely by users worldwide, including users far beyond the service area of the physical library. The Digital Initiative would raise the library's profile beyond the library's service region by showcasing the richness of the collection on the web.

Digitization as preservation: The Archives' collections include items that are rare, fragile, unique, or all three. Digitization thus would serve as a means of preserving the contents of these materials in the long term. Furthermore, access to these materials was often restricted, because the materials were in fragile condition. Online access to digital images of the materials would provide much greater access, and at the same time decrease the demand for physical handling of these materials.

Digitization as cost-saver: Contrary to Hughes (2004, 51), the Archives believed that enhanced online access to digital images would reduce traffic to the Archives Reading Room, thus shifting some of the archivists' time from providing reference to processing collections. The reduced use of the physical materials would also allow for more compact shelving. Finally, the ease of access to high-resolution images would improve efficiency of the Archives' fee-based digital imaging service (discussed in the next section).

Lessons: (1) Digitization programs can be born out of the most urgent and unexpected circumstances, and administrators need to be prepared to take over the program at any state of gestation; (2) Digitization program staff need to be aware of the parts of the library's mission that supports the digitization program; (3) Receiving continual support from the senior leadership is crucial; (4) The Digital Initiative brought many benefits to the library that went far beyond simply adding a web presence.

Digital Initiative Program Development

Activities Prior to the Digital Initiative

Electronic photographs database: The QBPL Archives had first experimented with digitization in the mid-1990s. The Division had digitized over 70,000 photographs and postcards, and created a database using the software ApplicationXtender by a content management software vendor EMC. The staff spent roughly 5 years on the digitization, and many years thereafter entered image descriptions. Images were digitized at a wide range of resolutions and workmanship. Early in the project, images were created at 72 dpi using a hand-held digital camera; later in the project, a scanner was acquired and images were scanned at 300 dpi. The digital images had different borders and shading, and some images were out of focus.

All the images were and continue to be organized in the software database ApplicationXtender with some accompanying descriptive metadata, including photographer, year, location, category, and image description. The database can still be searched at one public terminal in the reading room and at one other staff terminal. The search engine supports some advanced search options, but the interface only provides searches for the exact data fields, which creates some inconvenient consequences. For example, image descriptions span four 128-character fields, and searching one field does not automatically search any of the other three. As another example, when metadata are exported to spreadsheets, peculiarities with the data are preserved in the export: locations and categories follow in-house controlled vocabularies that are inconsistently applied; descriptions longer than 128 characters flow into the next field; and text strings are all in capital letters.

This database has been serving as the main discovery tool for the Archives' photograph and postcard collections. The item-level descriptions of images have proven to be a great asset.

The majority of library patrons have been able to navigate the interface and retrieve some images from the system. They have also been generally satisfied with seeing the images and descriptive metadata on the screen and seldom request to see the physical items.

Fee-based digital imaging service: The Archives offers a fee-based scanning service. After searching the image database, patrons can request high-resolution images scanned at 600 dpi. The Archives receives roughly 400 such requests annually, bringing in roughly $8,000 in revenue. However, the existing database cannot handle images in the TIFF format, the format in which these high-resolution images are scanned. So, without a viable way to add these images to the existing database, the newly scanned high-resolution images are deleted after each request.

"The authors reported that an elegant teaching method for student learning is to incorporate online tutorials into a lecture-based session followed by an exercise with evaluative worksheets."

Because of this fee-based service, the Archives had been equipped to support a rudimentary digital imaging operation. It already owned a high-end flat-bed scanner and image-editing software, and had staff members who are trained to maintain the equipment, use the equipment, and provide the service. But as we came to realize, what the library needed was a digital imaging program. As far as this fee-based service was concerned, a robust digital imaging program would be able to facilitate selection and retrieval of these images, save staff time for locating and scanning the physical items, and, hopefully, increase revenue. Although maximizing revenue was not the mission of the library, as we saw it, this service could potentially bring in much-welcome extra income to the Archives division.

<u>Lessons:</u> (1) A database design that adheres to open standards for its metadata and interface can increase its long-term viability as technologies change; (2) Remote access options through standard protocols (such as web access) should be seriously considered; (4) A digital imaging policy should ensure uniform image quality; (5) Program administrators should take an inventory of existing services and equipment as part of the program development process.

Program Mission:

The mission statement for the Digital Initiative (QBPL 2009b) borrowed some language from the Library's mission statement, but specifically referred to aspects that were unique to the program. (See Appendix A). The main objective was to digitize materials and make them available online, and the scope was all of the library's archival collections. Two key points in this statement had significant practical importance in terms of influencing the subsequent development of the program:

t

(a) The Digital Initiative was to serve as the digital repository for all the Library's unique and special collections. These might include existing and future collections, and might also encompass institutional records (because they are unique). The scope of the digital collection was later to be further codified in a Collecting Policy document.

(b) The Digital Initiative was to develop in three areas: database, metadata, and standards. Linking these three seemingly disparate areas was the unique feature that made the Digital Initiative a sustainable program, rather than a one-time project. Developing a database implied maintaining a formal database development life cycle; developing robust metadata would ensure proper item inventory and descriptions, and would prevent technological obsolescence; adhering to standards would ensure an open, widely compatible system that would be able to adapt to technological change at all levels, including changes in database technology, imaging standards, metadata standards, and web discovery tools.

According to the manager of the program, drafts of the Digital Initiative's mission statement were circulated and inputs were solicited from the library's senior leadership, especially from divisions that would eventually participate in digitization operations. After several rounds of discussions and revisions, the mission statement was approved by the Library Director.

Lessons: (1) The mission statement should mirror the library's mission, while including elements specific to the digitization program that ensure the program's long-term sustainability; (2) Serving as the institutional repository can be a monumental task, but can also help secure sustained funding to the program in the future as the repository becomes a significant part of the institution; (3) Program administrators should maintain constant communication with the senior leadership at this critical phase, garnering their participation and establishing their sense of ownership and responsibility to the program.

Feasibility Study

Even after the mission statement had been drafted, the Digital Initiative was only an idea on paper. The next step was to establish that the idea was actually feasible. However, performing a feasibility study for the Digital Initiative would have been moot, because it had already been established at this point, feasible or not. Nonetheless, this was an opportunity to analyze the work ahead. Since the scope of the program had been laid out, this analysis could delve into the triple constraints of project management: quality, cost, and time.

The two main deliverables outlined in the mission statement were the database and the metadata. The actual scope of the work entailed all of the collections in the QBPL archives (QBPL, 2011b):

- 36,000 monographs
- 2,500 cubic feet / 100 collections of manuscripts

- 4,500 maps and broadsides
- 105,000 photographs
- 425 feet of vertical files (roughly 2,500 files)
- 9,000 reels of microfilm (roughly 200 titles)

This list would require a total of roughly 150,000 discrete catalog records.

For the digital assets database, the library's team of system programmers and IT department was already maintaining an ILS with over 7 million items in its collections (QBPL, 2011a). For 150,000 digital items, the library could either purchase or develop a system and manage its growth.

Most monographs, serials, and vertical files had already been cataloged in MARC; all manuscript collections had a finding aid; most photographs already had descriptions written. As a library that was acquiring an average of $9.0 million worth of new items annually in the past three years (QBPL 2010, 2009a, 2008), the new cataloging for cartographic materials, the catalog revision and data encoding and conversion could be absorbed into existing workflow and accomplished over the long term.

In terms of the triple constraints, the quality of the digital assets database was closely tied to its specification dictated in the Digital Initiative's mission statement: adherence to open standards and access for "generations of researchers" (QBPL 2009b). Interestingly, time was not constrained, which implied that the quality of the work should take precedence over speed. The cost for launching this program would be minimal—it could even be launched at no additional cost since the library already had existing resources—scanning equipment, catalogers, systems staff, and IT staff—to absorb the additional workload. However, looking long term, additional staff might have to be added to administer the program and perform digitization job functions new to the library; additional equipment might have to be purchased for increasing speed and for imaging material formats not suitable for the existing flat-bed scanner.

<u>Lessons</u>: (1) In a proper feasibility study, a per-item cost estimate and a processing timeframe estimate would be useful; (2) Knowing the minimum resources required can smooth other departments' anxiety about competition for resources.

Administrative Design:

<u>Administrative home</u>: The Digital Initiative's objective was to digitize the holdings in the Archives. The feasibility study pointed out that a large amount of work would be devoted to creating metadata for the archival materials. The Digital Initiative should be administratively affiliated with either the cataloging department or with the Archives. Further consideration was given to minimizing handling the archival materials, so setting up office at the Archives was the best option. This administrative unit operated under the Digital Initiative umbrella, and was named the Digital Assets Management System (DAMS), physically sharing space in the technical services area in the Archives.

<u>Determining staffing level</u>: The feasibility study further revealed that collaboration across many departments was necessary—database development would involve systems programmers; data storage would involve the IT department; the user interface would involve the web design team; equipment purchases would involve the purchasing department; cataloging would involve the cataloging department. The library had two options: running the DAMS through committee, or assigning dedicated staff. The committee option would require gathering representatives from all participating departments, and the com-

mittee as a whole would supervise the operation of the digitization program. Regular meetings of this committee would maximize interdepartmental cooperation, but, while not requiring any additional staff, the program would not be the main focus of any one department. In contrast, a dedicated staff, even at a minimal level, could focus on handling all the unique functions of a digitization program and developing area expertise, while reaching out to other departments for collaboration. In fact, DAMS would benefit from having a manager who provided cohesion and leadership. The manager's job would be to represent DAMS internally and externally, facilitate interdepartmental collaboration, make day-to-day decisions, and manage all the unique functions of the program.

After weighing the two options, the library's senior leadership decided to assign a dedicated staff. The process of determining staffing level further identified two major job functions that did not exist in the library's structure at the time—metadata creation for digital objects and the actual digitizing. DAMS was fortunate enough to be able to acquire approval for all three positions (manager, metadata librarian, and digitization technologist), and filled these positions through internal transfers.

<u>Lessons</u>: (1) The Digital Initiative could have been run out of any division or could have existed as a separate division in the library. But since a digitization program is like setting up a separate library and involves bringing so many separate library functions together, it is best for the program to be a single, dedicated administrative entity so that ground-level, day-to-day decisions can be made quickly; (2) One additional advantage is that a dedicated staff can develop expertise over time, which can lead to sustained interest and strength of a digitization program (also see Hughes 2004, 96-110).

Budget Considerations:

Program budget: It was no coincidence that all three DAMS staff positions were filled by internal candidates, because internal transfers did not incur any additional cost in the library's over all staffing budget. So, other than a small supplies budget, the cost of setting up DAMS was minimal. Discussion on the cost for the actual digitization work is found in the following sections.

In-house vs. outsourcing. Hughes (2004, 93-97) weighs the advantages and disadvantages of contracting external vendors for digitization work. In DAMS's case, the decision was mainly guided by the library's funding mechanism for capital and operational expenses. An in-house digital imaging program required acquiring equipment, which would be purchased through the capital budget, whereas a digitization service contract with an external vendor would be paid through the operating budget. For an in-house operation, a book scanner, an overhead camera setup, the maintenance package cost roughly $100,000, which was miniscule compared to New York City's average annual capital procurement of $10.2 billion from 2008 to 2010 (NYCOMB 2007, 2008, 2009). However, allocating the same amount for an outside vendor in the library's annual operating budget, which averaged $127.2 million in the same three years (QBPL, 2008, 2009a, 2010), represented a much larger proportion, and would compete with many other library programs that provide vital services to the community (enough for one branch library to open on Saturdays for a year, for example).

Moreover, the outsourced portion covered only part of the digitization process. The library still would still need to spend resources on cataloging, hosting, and maintaining the digital assets. With the digitization equipment in-house, the library would gain greater flexibility over the pace of digitization, the design of the workflow, and the overall quality control of the process. In the absolute worst case scenario, where no operating budget was set aside for the program, the program could still potentially be run entirely by volunteers if the equipment were there. So, the DAMS manager initiated the procurement process with the city government.

Lessons: (1) Contracting an external vendor can speed up the process, but can also become very expensive in the long term; (2) "Developing policies for the worst-case scenario can help boost the imaging program's resilience to ever-fluctuating economic conditions."

Sustainability Considerations

After setting up the workspace and acquiring staff, the final part of developing the digitization program was to create policies and practices that would ensure its long-term sustainability as an organization.

Organizational Sustainability:

In order for DAMS to be sustainable, its organizational structure needed to be robust, yet flexible enough to adapt to changing needs and evolving technology, and its web presence needed to be aggressively promoted and updated. Internally, informal staff meetings kept all staff current on projects and allowed the manager to update other staff on external relations. The manager maintained regular contact with areas of the library that would become heavily involved with digitization work through monthly meetings with the heads of cataloging, systems, IT, web development, as well as administration.

Technological Sustainability:

Even though earlier analysis had shown that the demands of a new digital assets database could be met with existing IT infrastructure, DAMS would still make a number of specific demands to satisfy the unique needs of digital preservation. These included dedicated server space and shared drives, a high-capacity network among DAMS staff and between DAMS and the library's servers, as well as data backup and recovery procedures that met preservation standards. There was also a potential for an increased demand for bandwidth on the public web site after images went online. Although these demands were not unusual, the DAMS manager ensured that these requirements were met on an ongoing basis. The DAMS manager also discovered that the existence of a digitization program qualified the library for certain grant funding toward technology upgrades, which defrayed some of the cost.

Content Sustainability:

Sustaining the content would involve continual updates and securing rights for the digital assets. The DAMS manager drafted a collection policy to accompany the mission statement (QBPL 2009b). The collection policy specified what digital assets were to be included. Among the types of materials DAMS would collect were born-digital materials contributed by users, which included the public as well as the library's staff. This policy would allow for documenting the cultural heritage not only of the area, but also of the library itself, functioning as an institutional depository.

A copyright statement was formally adopted by DAMS (QBPL 2009b). This copyright statement was modeled after the one in use in the Archives, and had been approved by the legal department. The library held the rights to provide access for private study, while users were responsible for any other use of the materials. However, since the Archives held a large amount of public domain materials produced before 1923, further examination of materials under copyright has been postponed.

Lessons: (1) Keeping stakeholders informed of current project development can ensure efficient collaboration; (2) Digitization programs can serve as a grant funding source; (3) A broad, forward-looking collection policy can keep the digitization program at the forefront of library service; (4) The program can begin digitizing public domain materials while copyright issues are being resolved.

Conclusion

From establishment of the Digital Initiative to the launch of DAMS, the library staff examined what embarking on a digitization project might mean to the institution. This examination involved weighing the advantages and disadvantages of such a program, analyzing existing resources and activities, studying the program's feasibility, designing an administrative structure, and setting up sustainable practices. We learned that the new program should always strive to remain active, visible, and productive in the library. The program should set realistic milestones that are regularly met and communicated. The manager should find creative ways to incorporate the program's needs into other departments' existing operations. Acquiring other departments' buy-ins not only minimized disruption to their operations and minimized cost, but also gave their staff additional satisfaction from delivering a new product or service. Library administrators can extract key points from this narrative, discussion, and analysis when considering setting up a digitization program.

Reference List

Boock, Michael. 2008. "Organizing for Digitization at Oregon State University: a Case Study and Comparison with ARL Libraries." *Journal of Academic Librarianship* 34 (5): 445-451.

Federal Agencies Digitization Guidelines Initiative. 2010. "Guidelines." Last Updated: December 17.

Greenstein, Daniel, and Suzanne E. Thorin. 2002. *The Digital Library: A Biography*, 2nd ed. Washington, DC: Digital Library Federation Council on Library and Information Resources.

Hughes, Lorna M. 2004. *Digitizing Collections: Strategic Issues for the Information Manager*. London: Facet Publishing.

Kenney, Anne, R., and Oya Y. Rieger. 2000. *Moving Theory into Practice: Digital Imaging for Libraries and Archives.* Mountain View, VA: Research Libraries Group.

International Federation of Library Associations. 2002. "Guidelines for Digitization Projects for Collections and Holdings in the Public Domain, Particularly Those Held by Libraries and Archives." Unpublished.

Library of Congress. 2002. *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program, a Collaborative Initiative of the Library of Congress.* Washington, DC: Library of Congress.

Lopatin, Laurie. 2006. "Library Digitization projects, issues and guidelines: a survey of the literature." *Library Hi Tech* 24, no. 2: 273-89.

New York City Office of Management and Budget. 2009. "The City of New York, Adopted Budget, Fiscal Year 2010: Capital." New York City Office of Management and Budget, http://www.nyc.gov/html/omb/downloads/pdf/cb7_09.pdf

———. 2008. "The City of New York, Adopted Budget, Fiscal Year 2009: Capital." New York City Office of Management and Budget, http://www.nyc.gov/html/omb/downloads/pdf/cb7_08.pdf

———. 2007. "The City of New York, Adopted Budget, Fiscal Year 2008: Capital." New York City Office of Management and Budget, http://www.nyc.gov/html/omb/downloads/pdf/cb7_07.pdf

Queens Borough Public Library. 2011a. "About Us." Queens Borough Public Library, http://www.queenslibrary.org/?page_id=6

———. 2011b. "Archives at Queens Library." Queens Borough Public Library, http://www.queenslibrary.org/index.aspx?page_nm=CEL+Divisions&branch_id=CEL&pid=546

———. 2010. "Annual Report 2010: Financial Report." Queens Borough Public Library, http://www.queenslibrary.org/index.aspx?page_nm=Annual_Report10_FR

———. 2009a. "Annual Report 2009: Financial Report." Queens Borough Public Library, http://www.queenslibrary.org/index.aspx?page_nm=Annual_Report09_FR

———. 2009b. "Queens Library's Digital Initiative Mission Statement." Queens Borough Public Library.

———. 2008. "Annual Report 2008: Financial Report." Queens Borough Public Library, http://www.queenslibrary.org/index.aspx?page_nm=Annual_Report08_FR

———. 1991. "Mission Statement, adopted January 1991." Queens Borough Public Library, http://www.queenslibrary.org/index.aspx?page_nm=Mission+Statement

Sennema, Greg. 2004. "Developing a Digital Archive with Limited Resources." *OCLC Systems & Services: International Digital Library Perspectives* 20, no. 2: 76-81.

Smith, Abby. 2001. *Strategies for Building Digitized Collections*. Washington, DC: Council on Library & Information Resources. http://www.clir.org/pubs/reports/pub101/contents.html

Appendix A

Queens Library's Digital Initiative Mission Statement:

The Queens Library's Digital Initiative is a web-accessible repository for digital assets from the Library's collections. This unique and varied content documents Queens, Brooklyn, and suburban Long Island. The Initiative promotes the scholarship of Long Island by providing generations of researchers from around the world with expanded and enhanced access to this material using standard web browsers.

- The Initiative includes an intuitive web based discovery tool for customers to find the assets contained within the repository.
- The Initiative includes all the Library's unique and special collections.
- The Initiative is a participatory, interactive and collaborative repository for its customers.
- The Library utilizes the latest storage and web-based technologies and inter-operative descriptive standards for its content.
- The Initiative is comprised of database, metadata and digital format standards.

Collecting Policy

In an effort to fulfill the mission statement of the Queens Library's Digital Initiative the digital archives will contain the following material.

The digital archives will first and foremost consist of the Queens Library's unique special collections of the Long Island Division that document Long Island. This includes digitized monographs, serials, maps, photographs, newspapers, art work, broadsides, 3-dimensional objects and more.

The digital archives will also be the repository for customer's "born-digital" records documenting Queens, Brooklyn, Nassau and Suffolk. This digital archive will allow customers to submit their digital photographs, websites, blogs, newspapers and other digital media that document Long Island. The Library reserves the right to refuse the donation of materials.

Access, Reproductions and Intellectual property and donor restrictions

Furthering the mission of the Queens Library's Digital Initiative to promote the scholarship of Long Island, the digital archives will provide open and easy access to its contents, as well as reproductions of this content upon request of the customer. On occasion this open access will be mitigated by the United States law governing intellectual property rights and donors' restrictions detailed in the Deed of Gift.

The following is the digital archives' copyright statement. The Copyright law of the United States governs the reproduction of copyrighted material. The Long Island Division is authorized to provide reproductions of copyrighted material only if the reproduction is used for private study, scholarship or research. Be aware that responsibility for copyright clearance to reproduce the reproduction rests entirely with the user. Be aware the Long Island Division owns the physical object but does not necessarily own the copyright to the image.

In regards to access, the digital archives will provide access to all its content through the World Wide Web including the digital object and its corresponding catalog.

NYLA

JLAMS

A Peer-Reviewed Journal of the New York Library Association

Copyright, Digitization, & RFID

Fall 2011

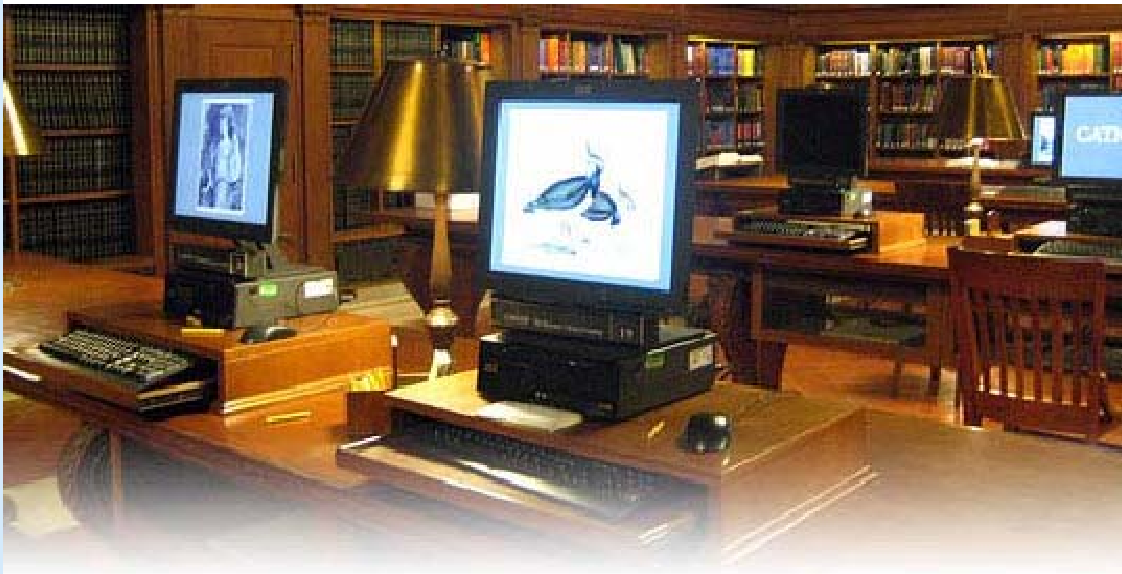
Vol. 8, No. 1

To submit an article read the JLAMS editorialpolicy on our websiteat:
http://www.nyla.org/index.php?page_id=813
Then send an eMail proposal to:
naylor@ColonieLibrary.org

New York Library Association
6021 State Farm Road
Guilderland, NY 10284
518-432-6952/800-252-NYLA email: info@nyla.org web:www.nyla.org

President
Marcia Eggleston
315-250-0352
megglest@nncsk12.org

Executive Director
Michael J. Borges director@nyla.org



CONTENTS

DIGITIZING EVERYTHING?

Part II: PILOTING METADATA CREATION

By Kimmy Szeto

Abstract: The Queens Borough Public Library launched the Digital Assets Management System in 2009 with the ultimate goal of digitizing all the holdings in the library's Archives. A pilot project was initiated in the course of designing the metadata creation policy and cataloging workflow. This paper discusses key policy and design elements such as imaging requirements, legacy data migration, metadata schemes, data formats, file naming, and controlled vocabulary, and presents sample data processing scripts, VRA Core 4.0 metadata records, and transformations to HTML and KML documents.

Introduction

The Queens Borough Public Library (QBPL) is nearing the public unveiling of a web site for its digitized archival collections dedicated to the history of Long Island. The driving force behind this effort is the Digital Assets Management System (DAMS), which was initiated by its current Project Manager, and was formally launched in 2009 after 3 years of planning (Szeto, 2011). I joined the digital program as its metadata librarian at that time, and participated in developing the program's metadata policy, workflow design, and implementation. In this role, I frequently drew ideas from current literature on digital imaging implementation, and at the same time invented solutions for site-specific problems. The detailed account of the metadata creation process for the digitization program's pilot project in this paper provides a practical perspective of our course of action.

Literature on Implementing Digitization Programs

Kenney and Rieger (2000) is a comprehensive guide to building and maintaining a digital assets collection; Hughes (2004) provides a balance of technical and strategic guidance. Best practices and technical guidelines for digitization of various formats can be found in a document published by the Federal Agencies Digitization Guidelines Initiative Still Image Working Group (FADGI, 2010). Case studies provide examples of how to develop practical solutions.

Kimmy Szeto is Assistant Librarian at the Maritime College, State University of New York. He heads cataloging and metadata creation at the Stephen B.. Luce Library and Archives, and manages the Federal Deposit Library Program at the library. His email is kszeto@sunymaritime.edu.

Novara (2010) discusses the digital imaging and archival practice at the University of Maryland, with emphasis on adapting workflow to the changing demands researchers make when it comes to digital technology. Schmidt, et al. (2011) detail the survey, evaluation, and planning carried out by the digital curation team at Michigan State University. Their findings address how one institution makes decisions on metadata schemas, controlled vocabularies, and digital storage solutions in response to various local practices. Fox (2008) focuses on descriptive metadata for digital images and carefully examines the VRA Core schema and its role in capturing metadata for cultural heritage materials. Colati and Colati (2011a, 2011b) impart many words of wisdom as they chronicle the effort of a fictitious academic librarian who accidentally and reluctantly found himself tasked with establishing a digital contents management program.

Deepening the level of detail that such sources have begun to provide, this paper focuses on solving metadata problems during the first implementation of a digitization program. Presenting this experience will help other libraries overcome some of their implementation hurdles.

Project Background

The Digital Assets Management System and the Archives

In 2006, the Queens Borough Public Library established its Digital Initiative to digitize the contents of its archives. After the initial feasibility study and organizational structuring, the Digital Assets Management System was set up as a subdivision of the Archives. By 2009, it had established its mission and collection development policy, had been staffed with 3 FTEs, and had had hardware and software procurement funding approved. The scope of this program was ambitious: it would digitize all the contents of the library's Archives and the library's institutional records, as well as collect all born digital contents emanating from the library. This presented a curation predicament similar to the one found in Schmidt, et al. (2011)—each material format in the Archives had its own set of metadata, and each individual agency produced its own electronic documents.

The holdings of the Archives consist of roughly 3,600 monographs, 2,500 cubic feet of manuscripts (in about 100 collections), 4,500 maps and broadsides, 105,000 photographs and postcards (in about 50 collections), 425 feet of vertical files (about 2,500 files), and 9,000 reels of microfilm (about 200 titles) (QBPL, 2011). All the monographs, vertical files, and titles on microfilm have MARC records and are searchable on the library's OPAC. In 2009, the cataloging department had just begun to catalog all the maps and broadsides systematically, also in MARC. Manuscripts had all been accessioned by librarians in the Archives, who had also created finding aids for each collection. These finding aids were in the process of being encoded in EAD at a steady pace by volunteers. Earlier in the mid-1990s, most photographs had been described at the item level and digitized at low resolution. Ever since then, these images have been available for searching and viewing at a computer terminal using the ApplicationXtender database software by EMC.

With the procurement process ongoing, it would take another 6 to 9 months before the new imaging system materialized; at the same time, the software company contracted for the digital contents management software was experiencing production delays. In the meantime, the digital imaging program, firmly established and fully staffed, found itself without the equipment to do the work. However, there was one area to explore. The Archives Division owned a high-resolution flat-bed scanner, and offered a fee-based scanning service, where library patrons could request high-resolution scans of images discovered with the aging ApplicationXtender.

We saw a convergence of factors: a high-quality scanner, an outdated database, and a need to improve access. It made sense to digitize photographs systematically using the existing flat-bed scanner, and delve into metadata creation—developing policies, creating workflow, and migrating legacy data—through a pilot project.

Existing Digitized Assets

A closer look into ApplicationXtender and the existing fee-based imaging service revealed troubling news. The images in the database were in JPEG format, ranging from 72-dpi photographs (of the photographs) taken by a hand-held digital camera, to 300-dpi scans with many defects. The accompanying descriptive metadata include photographer, year, location, an in-house subject, and a few lines of description. The search engine is symptomatic of search interfaces of its day. For example, the description field is broken into four separate 128-character text fields, and searches could only be performed on one field at a time. (The workaround, obviously, is to perform the same search four times.) Nevertheless, the 70,000 item-level descriptions were the strength of this system. The majority of patrons had learned to work with this interface, and they generally had been able to fulfill their research needs. The system had been effective enough to keep the fee-based imaging service going.

This imaging service received roughly 400 requests annually. When a request came in, one of the librarians would retrieve the print or negative for the requested photographs, and then the technician would scan at 400 dpi and turn over TIFF images on a CD. However, ApplicationXtender cannot handle the TIFF format, and, as a result, these images were simply being deleted.

This situation was similar to the “serious image management problem” described in Novara (2010), where new digital files were continuously being scanned, and the database was inaccessible and user-unfriendly. At the same time, long-term preservation strategies and metadata specifications were being developed while waiting for a new digital contents management system to be brought online in the near future.

We took several steps similar to those described in Novara (2010): setting imaging specifications and developing nationally compliant metadata policy. We also noted the difficulties Novara (2010) encountered with the lack of hierarchical and lateral relationship between items and collections, and focused on the relevant analysis of the structure of VRA Core schema in Fox (2008). We further took into account of the significance of OAI-PMH compliance described in Colati and Colati (2011b). We also examined the dual TIFF dark archive and JPEG access derivatives arrangement in Schmidt, et al. (2011). Finally, we studied the significance of collaborative outreach in Schmidt, et al. (2011) and did some of our own.

Preparing for the Pilot Project

Imaging requirements

We adopted imaging requirements directly from the best practices document published by the now-defunct Bibliographic Center for Research Collaborative Digital Program (CDP, 2008), with some additional details from American Memory’s *Technical Information* site (American Memory, n.d.), most of whose guidelines have since been incorporated into FADGI (2010), and from the Arts and Humanities Data Service *Guides to Good Practice* site (Dunning, 2008). These imaging requirements include technical specifications such as work environment, scanner performance, calibration, color management, image processing, as well as specific instructions for quality management.

We recognized that the decision with the widest impact would be the choice of file formats.[1] For the pilot project, as well as all subsequent imaging, we chose TIFF for the digital master and JPEG 2000 for the access derivative.

TIFF is the appropriate choice for several reasons. It is a widely-accepted, open standard, non-proprietary format, which ensures interoperability and should not present complications with the anticipated digital contents management software. The format also allows for a large number of device-independent color spaces, and

1. *A detailed comparison of many image formats can be found on page 67-68 in FADGI (2010).*

supports embedded technical metadata, which is a potential time-saver when scanners are configured to supply the information automatically.

JPEG 2000 is gaining increasing acceptance among software applications. In addition to data compression, which is essential for online delivery, it stands out because it can embed metadata and support multiple resolutions. The latter feature can streamline file management: one single file can support displays at multiple resolutions, from thumbnail to high-resolution.

Selecting a Metadata Scheme

The digital content management software would be responsible for storing the digital objects and their metadata in XML, as well as generating its own preservation and administrative metadata. So, we focused on the descriptive metadata creation process. The only material format in the Archives without an established descriptive metadata scheme was the photograph (everything else was in EAD or MARC, which could later be mapped to MARCXML or other schemas). We wanted to select an existing widely-accepted schema that could handle hierarchical relationships. This eliminated Dublin Core, and left MODS, EAD, and VRA Core 4.0. MODS, like MARC/MARCXML, supports relationships between records in its *relatedItem* element (MARC 76X-78X linking entries). This functionality is intended to establish bibliographic relationships between separately-cataloged items. For cataloging collections of photographs, these fields can take on a new role in describing item-collection, item-item, and negative-positive-print relationships. In contrast, EAD and VRA Core 4.0 support these relationships natively. However, EAD is too robust and complex compared to VRA Core 4.0's three hierarchies: collection, work, and item. VRA Core 4.0 also provides fuller support for other art-related fields. For these reasons, we selected VRA Core 4.0 as the metadata scheme for photographs.

Internal Outreach

Before the pilot project began, we reached out to the IT staff and the web development team. At QBPL, there is a strong tradition of each library division maintaining its own data on network shared drives, but digital images can quickly fill up available quota, and the frequent transfer of large files can affect network traffic performance. Furthermore, storage solutions for long-term preservation of a digital repository require significant investment and active management.[2] In our case, this infrastructure would be implemented over time, but it is never too early to notify the IT staff. In fact, the head of IT immediately instituted a more stringent backup schedule, created extra dedicated storage capacity, and ordered additional network capacity to the imaging lab in anticipation of the new imaging devices.

The web development team would be responsible for creating the public site for search, browse, and display. After coming to understand the structure of VRA Core, the web team would develop an interface with faceted search capabilities. However, the team would not develop an interim interface before the content management software was fully installed and tested.

Implementing the Pilot Project

Inventory and Legacy Data

The goal of the pilot project was to digitize and catalog 176 photographs taken by Hal B. Fullerton between 1880 and 1910, mostly on the topic of transportation—locomotives, railroads, train stations, and automobiles—in Kings, Queens, and Suffolk. This collection was chosen for several reasons—it was relatively small, the provenance was known, it was in the public domain (all photographs had been taken before 1923), the subject matter was narrow, and the collection consisted of a mix of media (prints, glass negatives, nitrate negatives, and interpositives) and sizes. The imaging technician had just scanned this collection, so

2. *Data storage recommendations can be found in Section VII of FADGI (2010).*

Table 1 . Excerpts from data exported from ApplicationXtender

CONTROL	SITE	DATE	SUBJECT	PHOTOGRAPHER	DESCRIPTION
HBF-9942	Queens Village (New York, N.Y.)		Long Island Railroad Company / Accidents	Fullerton, Hal B.	Long Island Railroad Company ACCIDENT – MEN LEANING AGAINST A FENCE LOOKING AT THE ACCIDENT – PRINT
HBF-9923	Long Island City (New York, N.Y.)	1899	Long Island Railroad Company / BOATS and SHIPS / DOCKS and PIERS	Fullerton, Hal B.	THE Long Island Railroad Company FERRY terminal at Long Island City (New York, N.Y.)
HBF-9920D	Queens Village (New York, N.Y.)	1898	Long Island Railroad Company / Accidents	Fullerton, Hal B.	clearing the SITE and REMOVING the WRECKAGE at Queens Village (New York, N.Y.), AFTER DERAILMENT of the Long Island Railroad Company TRIPLE-HEADER RUSSELL WEDGE PLOW and SUBSEQUENT FIRE / NOV. 28, 1898

both high- and low-resolution images were available. With the help of systems librarians, ApplicationXtender records for the collection were exported to a spreadsheet.

The legacy data presented quite a few challenges. Accession numbers, called Control Numbers, could not be reliably sorted, for example: HBF-93, HBF-846, HBF-1409B, and HBF-5376A-2. Some items were missing the date, which is a required element in VRA Core 4.0. All other text fields, including location, photographer, subject, and description, were all in capital letters. Upon further inspection, there were quite a number of misspellings and errors in these fields. (See Table 1 above.)

The capitalization of the location, photographer, and subject fields would only need a quick fix. However, the description fields were a different matter, even after concatenating the four fields. With capitalization problems, misspellings, and inaccuracies, it became clear that not all data fields could be directly migrated, and each photograph would need to be examined and cataloged individually. The next step, then, was to establish a cataloging policy and a workflow that maximized quality, accuracy and efficiency.

Cataloging and Metadata Policy

Our cataloging policy was derived from the Anglo-American Cataloguing Rules, second edition (AACR2). The minimum level of cataloging corresponded closely with the first level of description set forth in AACR2 1.0D1: title proper, first statement of responsibility, edition, material, publisher, date of publication, extent, notes, standard number, and terms of availability (Gorman & Winkler, 1998). A simple crosswalk was devel-

Table 2 . A preliminary crosswalk between AACR2, VRA Core 4.0 and simple Dublin Core.

AACR2	**VRA Core 4.0**	**Dublin Core**
Title proper	title	title
First statement of responsibility	agent	creator / contributor
Edition	stateEdition	title
Material	material	type
Publisher	rights / source	publisher
Date of publication	date	date
Extent	measurements	format
Note(s)	description / inscription / location	description / coverage
Standard number	refid (attribute)	identifier
Terms of availability	rights / href (attribute)	rights / identifier

oped between these AACR2 elements and VRA Core 4.0 (as well as to Dublin Core for future work on OAI-PMH compliance) as shown in Table 2 above. Elements applicable to photographs in terms of VRA Core 4.0 included *title*, *agent*, *material*, *date*, *measurements*, and *description/inscription/location*, and attributes *refid* and *href*. These were the metadata we would collect in the cataloging process.

One aspect AACR2 does adequately not cover is hierarchical relationships. VRA Core 4.0 is designed to distinguish and relate three hierarchical levels: *collection* (collections of art works), *work* (the actual art works), and *images* (visual reproductions of art works). The VRA 4.0 definition of *work* and *image* present some challenges. As defined in the schema's documentation, *work* is "a unique entity such as an object or event," and *image* is "a visual representation of a work in either whole or part" (VRA…Introduction, 2007). In the element description, *work* is further refined as "a built or created object," and *image* "a visual surrogate of such objects" (VRA…Element Description, 2007). However, the distinction between *work* and *image* as applied to photography is not as clear. This topic sparked long discussions with the library's cataloging staff. Is the photographic print a *work*? What about the negative? If the print is a *work* and the negative is an intermediary, is the negative, then, an *image*? But the negative is obviously not an image of the print. So, is it a related *work*? Can the negative and the print be separate *works*? The final verdict was to consider the *work* as a "created object" in the abstract sense: the photographer's vision of the capture of a particular moment (or moments) in a particular field (or fields) of vision. (The plurals were to include multiple exposures, continuous exposures, and stereoscopic photography.) Construed in this way, tangible forms of these visual captures, such as negatives and prints, were all considered *images*. This demarcation worked particularly well for the Archives' photograph collections because they were mainly documentary—photographs of buildings, railroad stations, locomotives, street scenes, etc.; the negatives and prints were not artistically produced—many were commercially manufactured and developed.

These definitions for *work* and *image* also worked well for catalogers, because catalogers could draw parallels to the cataloging concepts of manifestation and expression. This division also conveniently separated all the content description and subject analysis in *work* records from physical characteristics in *image* records. This separation would later influence workflow significantly.

Controlled Vocabulary

Other than the free-text photograph description and the notes fields, all other fields in VRA Core 4.0 descriptive metadata can be assigned to a controlled vocabulary. (The schema even provides the option for formatted dates.) In consultation with the cataloging department, we chose the Getty Research Institute's Art & Architecture Thesaurus (AAT) as the main vocabulary, given its appropriate scope and depth, and its ability to integrate with VRA Core. (For example, URI expressions were available for external data linking). In addition, Library of Congress Authorities would be used for elements that had direct bibliographic counterparts. These elements included names (personal and corporate), and subject headings (topical and geographic). Terms that did not have an authorized form would have one created.

A controlled vocabulary for geographic location presented more complexity. Most of the Archives' photographs were documented with street addresses. The Census Bureau's Master Address File (MAF) would be a great choice for a controlled vocabulary, especially given the potential to convert the MAF's address identifiers to uniform resource identifiers (URIs), which would enhance automated linking with any internet-based discovery system. However, neither the MAF nor the United States Postal Service database is available in list form. Using commercial products based on these databases or writing custom database interfaces was, unfortunately, beyond the scope of our program. Also, there were issues of general areas identified only by landmarks, obsolete street names, obsolete numbering, as well as non-addresses: for example, "Holtsville Station, near 985 Waverly Avenue, Holtsville, NY 11742," where the train station was razed in 1962; "Buhrman's Store, Bayside, near Alley Pond, now near West Alley Road and East Hampton Boulevard, Oakland Garden, NY 11362," where neither the store, the pond, nor the road where the store once stood exists today.

Even though we could not resolve the address problem when we were working on the pilot project, we recorded full addresses with structured punctuation and keywords such as "and" (to denote intersection), "near," and "now." This information would be sufficiently formatted for machines to parse, should a controlled vocabulary be instituted in the future.

With geographic data processing in mind, we decided to include longitude and latitude information as well. These geographic coordinates could be obtained through an online tool such as Google Maps, which catalogers would most likely be using anyway to verify addresses during the cataloging process. VRA Core 4.0 would be able to accommodate these data through the *extent* attribute.

We made several other decisions about controlled vocabulary: using inches for dimensions, since the majority of the photographic papers were 3"x5", 4"x6" or 8"x10", and conversion to centimeters could be automated; setting the level of granularity to distinguishing between black-and-white and color prints and between negatives from interpositives; and using "digital" as the material type for scanned images.

Data Dictionary

A data dictionary is an essential reference for ensuring uniform use of schema elements, especially when cataloging responsibilities will be distributed to a team of librarians. The one we developed was derived from the VRA Core 4.0 data dictionary (VRA…Element Descriptions, 2007) with additional information specific to DAMS. Controlled vocabularies were specified for their respective elements; frequently-used elements and attributes were highlighted; a commentary area was added to explain some of the rationales and intended scope for each element, sub-element, and attribute. We hoped that all this information would ensure uniformity and enable sound judgments.

Based on this data dictionary, a full crosswalk from VRA Core 4.0 to simple Dublin Core was also developed for future OAI-PMH compliance. DAMS was among the first to develop this crosswalk, but it is now included in *VRA Core 4.0 Element Description* (2007). An excerpt is shown in Table 3 on the next page.

Workflow and Automation

As we created descriptive metadata for the pilot project, the workflow and data entry interface for metadata creation were developed along the way. First, a number of global regular expression search and replace routines were built to migrate the legacy data from ApplicationXtender to useful form. They were applied to the Control Number, photographer, date, and subject headings in order to correct capitalization and formatting. Similar routines were applied to output names and subject headings in LC-authorized form. This method was able to correct the majority of the entries, which would save catalogers from having to correct each entry manually. We tested the method during the pilot project. In the future, these routines can be re-applied to all other legacy data.

The photograph description fields contained so many inaccuracies and errors that they were not reusable. However, these descriptions had originally been supplied by expert historians who identified the time, location, event, people, and objects in many photographs. Therefore, we decided to transfer the legacy description field directly to the new records but only make it visible to catalogers. Catalogers, then, would create a new description and supply a unique title.

There were two physical characteristics of each photo that we recorded. First, the material of the negatives and the prints; second, the dimensions. These data were recorded in the *image* records.

When we looked at the cataloging process as a whole, we found a distinct difference between the expertise for creating a *work* record and an *image* record. In the *work* record, we needed to inspect name and subject headings for the LC-authorized form, to create a free text description, and to supply a title. This would be a process

Table 3 . Excerpts from the DAMS Photographs Data Dictionary.

Queens Borough Public Library Archives - Photograph Collections - Data Dictionary

VRA Core 4.0 Re-stricted Section	XML Wrapper Name	XML Ele-ment Name	XML Sub-element Name	XML Attri-bute Name	Description	Data Type	Au-thority (local de-cisions)	Additional Infor-mation	Correspo nding Dublin Core Element	Man-datory?	Uni-que?
AGENT	agentSet	--	--	--	Contains elements that describe the names, appellations, or other identifiers assigned to an individual, group, or corporate body that has contributed to the design, creation, production, manufacture, or alteration of the work or image.	Element container	none	--		No	Yes
AGENT	agentSet	display	--	--	A free text note about AGENT.	Free text	none	--		No	Yes
AGENT	agentSet	agent	--	--	Contains elements that describe a single agent.	Element container	none	If there is more than one agent, the attribute *extent* is used.	CREATOR; CONTRIBUTOR	No	No
AGENT	agentSet	agent	--	extent	Qualification of the <role> subelement.	Free text	LCSH	Describes the part of the works or images that the agent is associated with.		No	Yes
AGENT	agentSet	agent	--	vocab	Describes the controlled vocabulary source from which *extent* is recorded.	Free text	none	Example: LCSH		No	Yes

Table 3 Part 1 continued

AGENT	agentSet	agent	--	vocab	Describes the controlled vocabulary source from which *extent* is recorded.	Free text	none	Example: LCSH		No	Yes
AGENT	agentSet	agent	name	--	Name, appellation, or other identifier assigned to an individual, group, or corporate body that has contributed to the design, creation, production, manufacture, or alteration of the work or image.	Free text	LC or LC style	--		No	Yes
AGENT	agentSet	agent	name	type	Qualification of <name>.	Free text	VRA 4.0 Restricted schema	Data values: "personal"; "corporate"; "family"; "other"		Yes	Yes
AGENT	agentSet	agent	culture	--	Name of the culture, people (ethnonym), or adjectival form of a country name from which the Collection, Work or Image originates, or the cultural context with which the Collection, Work, or Image is associated.	Free text	none	Currently unused. Use only for describing the AGENT. Use CULTURAL CONTEXT when describing a Collection, Work, or Image.	COVERAGE	No	Yes
AGENT	agentSet	agent	dates	--	Contains elements for date or range of dates associated with <name>.	Element container	none	Use only when referring to the Agent.	DATE; COVERAGE	No	Yes
AGENT	agentSet	agent	dates/earliestDate	--	Date of birth of an individual or head of family; date of founding of a corporation.	Free text	none	Format: YYYY, YYYY-MM, or YYYY-MM-DD.		Yes	Yes
AGENT	agentSet	agent	dates/latestDate	--	Date of death of an individual or the last member of a family; date of closing of a corporation.	Free text	none	Format: YYYY, YYYY-MM, YYYY-MM-DD, or *present*.		Yes	Yes
AGENT	agentSet	agent	dates/earliestDate or latestDate	circa	Use for approximate dates.	Free text	VRA 4.0 Restricted schema	Data values: "true"; "false"		No	Yes
AGENT	agentSet	agent	role	--	Qualification of <name>.	Free text	AAT	Use the singular form.		No	Yes

familiar to catalogers. For the *image* record, the tasks of selecting the appropriate AAT material term and supplying dimensions (or dpi for the digital surrogate) could be performed by paraprofessionals. A time-saving strategy would be to have the imaging technician perform those two tasks at the time of scanning. In fact, the cataloger's and the technician's parts need not take place in any particular order. With the help of the old descriptions, catalogers could even proceed with the digital surrogate alone. They would need to examine the actual photographs in only a small number of cases. This strategy would work as long as the *work* and *image* records could be merged. The possibility of catalogers and technicians working in parallel opened up many possibilities for the workflow, and influenced the automation and data input strategies.

Record Identifiers and File Naming

File names can serve as convenient identifiers for processing and retrieval.[3] However, they cannot serve as long-term metadata records, because their construction is limited by the file system, and files can be renamed. The format of the file names should also strike a balance between being useful and being too long and complex. In the case of the Archives' photographs, we wanted file names to reflect each photograph's Control Number, which is a unique accession number. Many of these numbers, however, needed leading zeros added so that they would be sortable. Since none of the photograph collections had more than 10,000 items, four digits were used. We added the leading zeros using a regular expression search and replace directly on the legacy data and replaced all file names using a freeware utility called the Bulk Rename Utility. Then we gave all the files a uniform three-letter extension that reflected the file format—.tif and .jpg. An added benefit to this file naming convention is that it facilitates any future automatic processing—Control Numbers can be extracted from file names, and file names can be constructed by extracting the Control Number from the record.

The VRA Core 4.0 schema requires an XML identifier for each record. In our case, conveniently, the newly formatted Control Numbers could be used as the record ID for *work* records, and the file name could be used in *image* records.

Data Entry

One other question that arose during the manual cataloging process was the data entry method. The metadata librarian began exclusively in an XML editor, but it would not be cost-effective to train every cataloging staff in the principles of XML and working directly with VRA Core 4.0 in its native XML form. The solution was to create data entry forms. After a few trials, neither Microsoft Word's field coding nor Excel's text export function resulted in a satisfactory data entry environment or output. Microsoft Word supports XML natively, and can handle some conditional situations (such as handling a variable number of subject fields), but it only checks for well-formedness, not for validity of the document. XML tags cannot be protected in Word, and it is quite cumbersome to work with tags if they are accidentally altered or erased. As for Excel, the usual method is to set up a spreadsheet with appropriate input fields while protecting the tags in surrounding fields, and then to export the spreadsheet as a text file. However, Excel's text export adds spaces between cells and unwanted quotation marks around text fields.

In our context, although these problems were not insurmountable, we had another goal, which was to maximize efficiency. VRA Core 4.0 presented a challenge, because the schema was designed to hold machine-readable data as well as human-readable expressions of the same data. For example, for a 5-inch by 7-inch black-and-white print, the numbers 5 and 7, and the AAT term "black-and-white print (photograph)" would appear in the *measurements* and *material* element sets. Then the *display* tags would read "5 x 7 inches" and "black-and-white print" as shown in Figure 1 below. Similar situations occurred with the date, name, and description fields. Entering *display* tags was redundant, and would cost a considerable amount of time and increase errors. Ideally, the *display* tags should be automatically generated based on the data. However, this kind of text generation is not possible in Word or in Excel.

3. "File names and organization of files in system directories comprise structural metadata in its barest form." (FADGI, 2010). Approaches to file naming naming are discussed in detail in the document.

Figure 1. Sample descriptive metadata and their expression in VRA Core 4.0

Control_Number	How many?	Width	Height	Material
HBF-0092	1	5	7	black-and-white negative
HBF-0092	1	4	5	interpositive
HBF-0092	2	8	10	black-and-white print (photograph)

```
<materialSet>
        <display>glass plate negative; interpositive; 2 prints; digital image</display>
        <material vocab="AAT" type="medium" refid="300128343">black-and-white negative</material>
        <material vocab="AAT" type="medium" refid="300137299">interpositive</material>
        <material vocab="AAT" type="medium" refid="300128349">black-and-white print (photograph)</material>
        <material vocab="AAT" type="medium" refid="300215302">digital image</material>
</materialSet>
<measurementsSet>
        <display>5 x 7 inches (negative); 4 x 5 inches (interpositive); 8 x 10 inches (2 prints); 400 dpi (digital image)</display>
        <measurements type="height" unit="inches" extent="negative">5</measurements>
        <measurements type="width" unit="inches" extent="negative">7</measurements>
        <measurements type="height" unit="inches" extent="interpositive">4</measurements>
        <measurements type="width" unit="inches" extent="interpositive">5</measurements>
        <measurements type="height" unit="inches" extent="print">8</measurements>
        <measurements type="width" unit="inches" extent="print">10</measurements>
        <measurements type="resolution" unit="dpi" extent="digital image">400</measurements>
</measurementsSet>
```

With that in mind, we turned our focus to developing text processing scripts while finding the most efficient interfaces for collecting only the data. For our imaging technician, the solution was straightforward. The inputs were width, height, and material. We set up a spreadsheet with these three fields, and the technician filled them out while scanning the item. We named it the Inventory Form, which was essentially the spreadsheet shown in Figure 1 above. For the catalogers, a similar spreadsheet proved to be too wide, due to the number of free text fields, and was difficult to see and navigate on the screen. We found another solution in Microsoft Access. We populated the same spreadsheet with an Access Form object that had a much cleaner and more intuitive interface. We named this interface the Descriptive Metadata Entry Form (Figure 2 on the next page) and the underlying spreadsheet the Descriptive Metadata Table. With this set up, controlled vocabularies could be directly linked to authorized fields via drop-down menus. This sped up input and reduced errors, and, it worked especially well for the pilot project, because there was only a single photographer and the collection covered a limited number of subject headings.

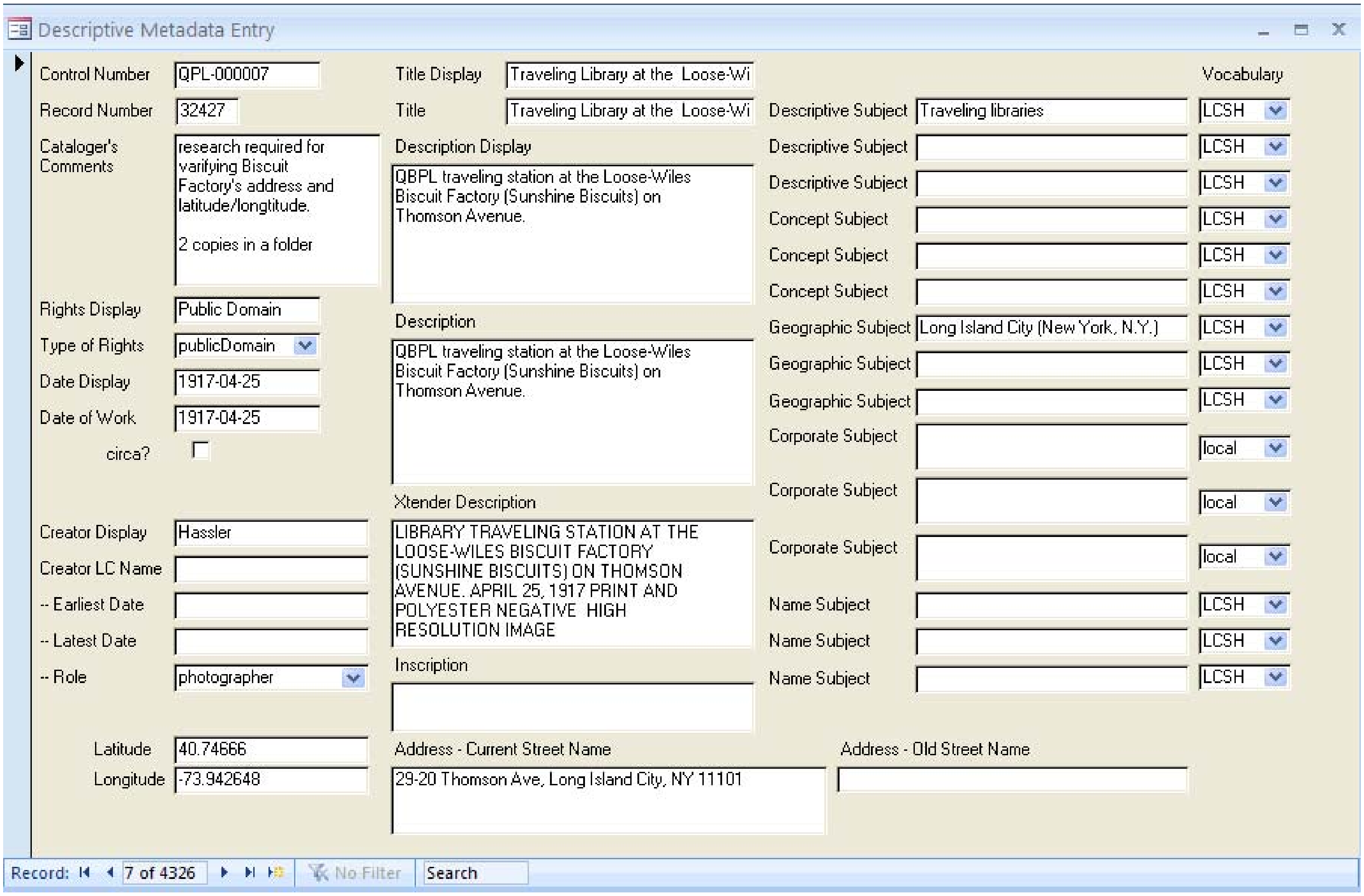
These two interfaces—the Inventory Form and the Descriptive Metadata Entry Form—served to collect the underlying descriptive metadata that would eventually be assembled into VRA Core records.

Data Processing and Metadata Creation

Descriptive metadata were assembled in two steps. The Descriptive Metadata Table was processed first to create valid VRA Core 4.0 records. These provisional records already had all necessary identifiers, internal relationships, and external links, thanks to the file naming convention, and could be ready for immediate public use. Then we inserted data from the Inventory Form to complete these records.

Populating metadata records with data can be done in many ways. We opted for Microsoft Word's mail merge function for processing the Descriptive Metadata Table, because the librarians had enough familiarity with Word that they could potentially run the process on their own in the future. During the mail merge, data were extracted from the Descriptive Metadata Table and inserted between opening and closing tags. Conditional formulas made up most of the merge codes, because whether new tags would be created depended on whether

Figure 2. Screen shot of the *Descriptive Metadata Entry Form*.



the data field was blank. A few attributes were also dependent on the data. For example, *circa* was evaluated based on the “yes” or “no” value in the “circa” column. An excerpt of this merge document is shown in Figure 3.

Figure 3. Excerpts from the Microsoft Word merge document that assembled data from the Descriptive Metadata Table and created provisional VRA Core 4.0 records.

```
<agentSet>
    <display>{ IF Creator="" "Unknown" Creator }</display>
    <notes> -- This is record no. { MERGEFIELD Control_Number }-- </notes>
    <agent extent="photographs" vocab="LCSH">
        <name type="personal">{ IF Creator_LC_Name="" "Unknown" Creator_LC_Name }</name>
        <dates type="life">
            <earliestDate { IF Circa="Yes" "circa=true" "" }>{ IF Born<>"" Born "Unknown" }11111111</earliestDate>
            <latestDate { IF Circa="Yes" "circa=true" "" }>{ IF Born<>"" Born "Unknown" }</latestDate>
        </dates>
        <role vocab="AAT" refid="{ IF Role="Photographer" "300025687" "" }">{ IF Role<>"" Role "" }</role>
    </agent>
</agentSet>
…

<subjectSet>
    <subject vocab="LCSH">
        { IF Subject_Descriptive1<>"" "<term type="descriptiveTopic" vocab="LCSH" { MERGEFIELD Subject_Descriptive1 }</term>" ""}
```

```
{ IF Subject_Descriptive2<>"" "<term type="descriptiveTopic" vocab="LCSH" { MERGEFIELD Subject_Descriptive2 }</term>" ""}
{ IF Subject_Descriptive3<>"" "<term type="descriptiveTopic" vocab="LCSH" { MERGEFIELD Subject_Descriptive3 }</term>" ""}
```

[…and so on]

For the 176 records of the Hal B. Fullerton Photographs, this merge process took about 5 seconds. Once these provisional records were created, we could then insert inventory data as they became available. The process of traversing and inserting XML tags was beyond the capability of mail merge, and a text processing script was the only option. We chose Python as the language, because it has built-in functions for hierarchical data structures like XML. The text processing script read the Inventory Form and inserted data into the appropriate tags. The script also examined the numeric data and controlled vocabulary, and composed the free text descriptions in the *display* elements for onscreen presentation. Returning to Figure 1, the script first inserted the data: width ("7," "5," and "10"), height ("5," "4," and "8"), and dpi ("400") in *measurements*, and the AAT terms ("black-and-white negative," "interpositive," "black-and-white print (photograph)," and "digital image") in *material*. Then, based on each value in *material*, the script filled the *refid* attribute with AAT's unique numeric identifier, and linked *measurements* to *material* through the *extent* attribute. Then, based on the three values in the "How many?" column, *measurements*, and *materials*, the script composed the phrases in the two *display* tags. Furthermore, the script also inserted a *depictedIn* relation in the *work* record to show the materials represented in the *image* record. Figure 5 on page 45 shows two examples of full records created by this process; the full workflow diagram is shown in Figure 4 on the next page.

Additional Considerations

Finally, we wanted to ensure the sustainability of the metadata, and to be prepared for worst-case scenarios. What if the content management software company were to go out of business? What if we were to lose funding for technical support and the software malfunctioned? We wanted to have the capability to set up a web site based on existing metadata and files alone without the content management software. This had already been accomplished by basing file names and basing XML identifiers on Control Numbers, and having external links automatically generated in the metadata creation process. These records and files were designed to contain sufficient information such that all search, retrieval, and online presentation functions could be performed by drawing information from the XML records alone, processing script such as Python scripts or XSLT (Extensible Stylesheet Language Transformations).

We performed two experiments to test search and retrieval operations on the Hal B. Fullerton records. In the first experiment, we ran our records through the XSLT stylesheet used for showing VRA examples on the Visual Resources Associations web site. The browser displayed the resulting HTML as expected. The second experiment simulated a search and filter by location. We wrote a Python script to parse the addresses and geographic coordinates from the metadata records, and then search for sets of photographs that were taken at the same location and group them together. To visualize this, we used the Python script to export the results to an XML file in the Keyhole Markup Language (KML), a schema used for displaying data in an earth browser such as Google Maps. The script processing took about 15 seconds, and the resulting display is shown in Figure 6 on page 47.

Conclusion

This pilot project served successfully as a testing ground and enabled us to develop cataloging policy documents, a metadata creation workflow, and useful tools and scripts. The experience will serve as the foundation for metadata creation at the production scale, and as the basis for future changes and improvements to the digi-

Figure 4. Flow chart showing metadata creation workflow for archival photographs.

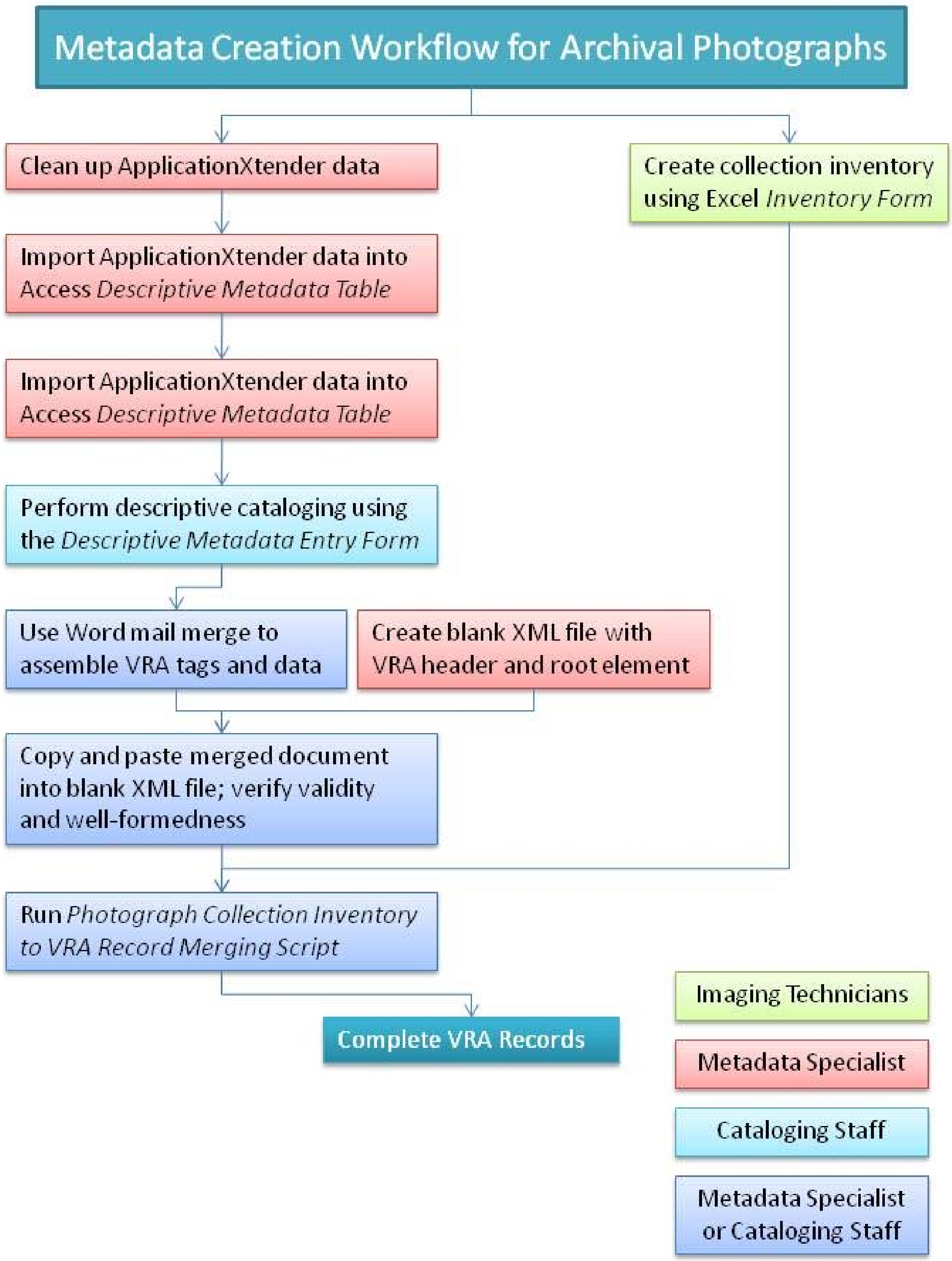


Figure 5. Two sample VRA Core 4.0 records.

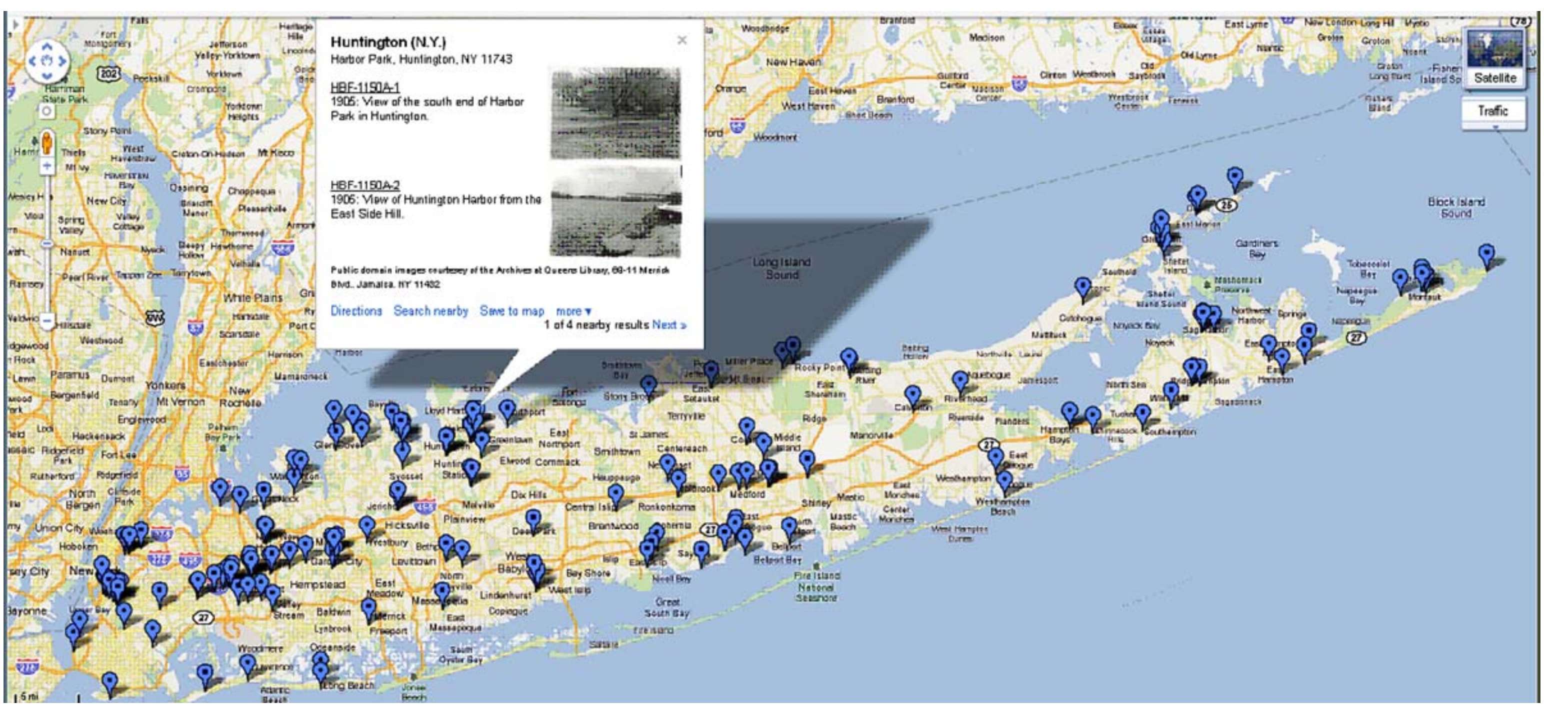
```
<work id="HBF-9920D" source="Hal B. Fullerton Photographs">
   <dateSet>
      <display>1898</display>
      <notes>---------- This is record no. HBF-9920D ----------</notes>
      <date type="creation">
         <earliestDate>1898</earliestDate>
         <latestDate>1898</latestDate>
      </date>
   </dateSet>
   <descriptionSet>
      <display>Removal of wreckage following derailment of Long Island Rail Road Company a 4-4-0
steam locomotive that was pushing a triple-header Russell wedge plow at Queens Village, November
28, 1898.</display>
      <description source="Xtender">clearing the site and removing the wreckage at queens village
(new York, n.y.), after derailment of the Long Island Rail Road Company triple-header russell
wedge plow and  subsequent fire / nov. 28, 1898 / print purchased from ron ziel.</description>
      <description source="KS">Removal of wreckage following derailment of Long Island Rail Road
Company a 4-4-0 steam locomotive that was pushing a triple-header Russell wedge plow at Queens
Village, November 28, 1898.</description>
   </descriptionSet>
   <locationSet>
      <location type="creation">
         <name type="geographic">Queens Village, NY 11429 (40.717703,-73.73597)</name>
      </location>
   </locationSet>
   <relationSet>
      <relation href="HBF_collection.xml" refid="HBF-9920D" type="partOf">Hal B. Fullerton
Photographs</relation>
      <relation relids="HBF-9920D.tif" href="URI of Image" refid="HBF-9920D"
type="depictedIn">negative; interpositive; print; digital image</relation>
   </relationSet>
   <rightsSet>
      <display>Public domain</display>
      <rights type="publicDomain"/>
   </rightsSet>
   <subjectSet>
      <subject vocab="LCSH">
         <term type="descriptiveTopic" vocab="LCSH">Railroad snowplows</term>
         <term type="descriptiveTopic" vocab="LCSH">Steam locomotives</term>
         <term type="conceptTopic" vocab="LCSH">Railroad accidents</term>
         <term type="geographicPlace" vocab="LCSH">Queens Village (New York, N.Y.)</term>
         <term type="corporateName" vocab="LCSH">Long Island Railroad Company</term>
      </subject>
   </subjectSet>
   <titleSet>
      <title source="KS">Wreckage of a Snow Plow Train</title>
   </titleSet>
   <worktypeSet>
      <worktype vocab="AAT" refid="300046300">photographs</worktype>
   </worktypeSet>
   <image id="HBF-9920D.tif" refid="X_28667" source="Hal B. Fullerton Photographs HBF-9920D">
      <materialSet>
         <display>glass plate negative; interpositive; 2 prints; digital image</display>
         <material vocab="AAT" type="medium" refid="300128343">black-and-white
negative</material>
         <material vocab="AAT" type="medium" refid="300137299">interpositive</material>
         <material vocab="AAT" type="medium" refid="300128349">black-and-white print
(photograph)</material>
         <material vocab="AAT" type="medium" refid="300215302">digital image</material>
      </materialSet>
      <measurementsSet>
         <display>5 x 7 inches (negative); 4 x 5 inches (interpositive); 8 x 10 inches (print);
400 dpi (digital image)</display>
         <measurements type="height" unit="inches" extent="negative">5</measurements>
         <measurements type="width" unit="inches" extent="negative">7</measurements>
```

```
      <measurements type="height" unit="inches" extent="interpositive">4</measurements>
      <measurements type="width" unit="inches" extent="interpositive">5</measurements>
      <measurements type="height" unit="inches" extent="print">8</measurements>
      <measurements type="width" unit="inches" extent="print">10</measurements>
      <measurements type="resolution" unit="dpi" extent="digital image">400</measurements>
    </measurementsSet>
    <relationSet>
      <relation relids="HBF-9920D" type="imageOf">Hal B. Fullerton Photographs HBF-
9920D</relation>
    </relationSet>
    <rightsSet>
      <display>Public domain</display>
      <rights type="publicDomain"/>
    </rightsSet>
    <sourceSet>
      <source>
        <name type="electronic">Hal B. Fullerton Photographs HBF-9920D</name>
        <refid type="URI">http://www.queenslibrary.org</refid>
      </source>
    </sourceSet>
  </image>
</work>

<work id="HBF-9970" source="Hal B. Fullerton Photographs">
  <dateSet>
    <display>ca. 1900</display>
    <notes>---------- This is record no. HBF-9970 ----------</notes>
    <date type="creation">
      <earliestDate circa="true">1900</earliestDate>
      <latestDate circa="true">1900</latestDate>
    </date>
  </dateSet>
  <descriptionSet>
    <display>LIRR ferry terminal, coal chute and power station from the East River.</display>
    <description source="KS">LIRR ferry terminal, coal chute and power station from the East
River.</description>
  </descriptionSet>
  <locationSet>
    <location type="creation">
      <name type="geographic">Long Island Rail Road Ferry Terminal, West 2nd Street and East
River, now near Borden Avenue and 2nd Street, Long Island City, NY 11101 (40.741822,-
73.961307)</name>
    </location>
  </locationSet>
  <relationSet>
    <relation relids="F012" href="HBF_collection.xml" refid="HBF-9970" type="partOf">Hal B.
Fullerton Photographs</relation>
    <relation relids="HBF-9970.tif" href="URI of Image" refid="HBF-9970"
type="depictedIn">negative; digital image</relation>
  </relationSet>
  <rightsSet>
    <display>Public domain</display>
    <rights type="publicDomain"/>
  </rightsSet>
  <subjectSet>
    <subject vocab="LCSH">
      <term type="descriptiveTopic" vocab="LCSH">Power-plants</term>
      <term type="descriptiveTopic" vocab="LCSH">Waterfronts</term>
      <term type="conceptTopic" vocab="LCSH">East River (N.Y.)</term>
      <term type="geographicPlace" vocab="LCSH">Long Island City (New York, N.Y.)</term>
      <term type="corporateName" vocab="LCSH">Long Island Railroad Company</term>
    </subject>
  </subjectSet>
  <titleSet>
    <title source="KS">Coal Chute and Power Station</title>
  </titleSet>
  <worktypeSet>
    <worktype vocab="AAT" refid="300046300">photographs</worktype>
```

```
    </worktypeSet>
    <image id="HBF-9970.tif" source="Hal B. Fullerton Photographs HBF-9970">
       <materialSet>
          <display>glass plate negative; digital image</display>
          <material  vocab="AAT"  type="medium"  refid="300128343">black-and-white
negative</material>
          <material vocab="AAT" type="medium" refid="300215302">digital image</material>
       </materialSet>
       <measurementsSet>
          <display>5 x 7 inches (negative); 400 dpi (digital image)</display>
          <measurements type="height" unit="inches" extent="negative">5</measurements>
          <measurements type="width" unit="inches" extent="negative">7</measurements>
          <measurements type="resolution" unit="dpi" extent="digital image">400</measurements>
       </measurementsSet>
       <relationSet>
          <relation  relids="HBF-9970"  type="imageOf">Hal  B.  Fullerton  Photographs  HBF-
9970</relation>
       </relationSet>
       <rightsSet>
          <display>Public domain</display>
          <rights type="publicDomain"/>
       </rightsSet>
       <sourceSet>
          <source>
             <name type="electronic">Hal B. Fullerton Photographs HBF-9970</name>
             <refid type="URI">http://www.queenslibrary.org</refid>
          </source>
       </sourceSet>
    </image>
</work>
```


Figure 6: KML output from the Python script processing of the VRA Core 4.0 metadata record displayed in Google Maps, showing the geographic distribution of the Hal B. Fullerton Photographs: blue pegs represent locations where photographs were taken; pop-up balloons display thumbnail images and descriptive metadata.



tization program. We look forward to working with content management software and with new imaging equipment, as well as future development of the web interface and further integration with geographic information systems.

Bibliography

American Memory. (n.d.) Technical information. Retrieved from http://memory.loc.gov/ammem/about/techIn.html

CDP Digital Imaging Best Practices Working Group. (2008). *BCR's CDP digital imaging best practices version 2.0*. Aurora, Colorado: Bibliographic Center for Research.

Colati, J. B., & Colati, G. C. (2011a). Road tripping down the digital preservation highway, part I: Hitting the road. *Journal of Electronic Resources Librarianship, 23*, 60-67.

Colati, J. B., & Colati, G. C. (2011b). Road tripping down the digital preservation highway, part II: Road signs, billboards, and rest stops along the way. *Journal of Electronic Resources Librarianship, 23*, 165-173.

Dunning, A. (2008). AHDS Guides to Good Practice. Retrieved from http://www.ahds.ac.uk/creating/guides/index.htm

Federal Agencies Digitization Guidelines Initiative Still Image Working Group. (2010). *Technical guidelines for digitizing cultural heritage materials: Creation of raster image master files.* Washington, DC: Federal Agencies Digitization Guidelines Initiative.

Fox, R. (2008). Culture byte by byte. *OCLC Systems & Services, 24*(4), 192-198.

Gorman, M., & Winkler, P. W. (Eds.) (1988). *Anglo-American Cataloguing Rules* (2nd ed., 1998 rev.). Chicago, IL: American Library Association.

Hughes, L. M. (2004). *Digitizing collections: Strategic issues for the Information Manager*. London: Facet Publishing.

Kenney, A. R., & Rieger, O. Y. (2000). *Moving theory into practice: Digital imaging for libraries and archives.* Mountain View, VA: Research Libraries Group.

Novara, E. A. (2010). Digitization and researcher demand: Digital imaging workflows at the University of Maryland Libraries. *OCLC Systems & Services, 26*(3), 166-176.

Queens Borough Public Library. (2011). Archives at Queens Library. Retrieved from http://www.queenslibrary.org/index.aspx?page_nm=CEL+Divisions&branch_id=CEL&pid=546

Schmidt, L, Ghering C., & Nicholson, S. (2011). Notes on operations: digital curation planning at Michigan State University. *Library Resources & Technical Services, 55*(2), 104-118.

Szeto, K. (2011). "Digitizing Everything?: The Launch of the Digital Imaging Program at Queens Library." *Journal of the Library Administration and Management Section of the New York Library Association, 7*(2), 5-15.

VRA Core 4.0 Introduction. (2007) Retrieved from http://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf

VRA Core 4.0 Element Description. (2007) Retrieved from http://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf

Paper 2

Positioning library data for the Semantic Web: recent developments in resource description

Journal of Web Librarianship, 7(3), pp. 305–321

DOI: 10.1080/19322909.2013.802584

Journal of Web Librarianship, 7:305–321, 2013
Published with license by Taylor & Francis
ISSN: 1932-2909 print / 1932-2917 online
DOI: 10.1080/19322909.2013.802584

Routledge
Taylor & Francis Group

COMMUNICATIONS

Positioning Library Data for the Semantic Web: Recent Developments in Resource Description

KIMMY SZETO
Newman Library, Baruch College, City University of New York, New York, USA

Recent developments in resource description standards and technologies have aimed at moving cataloging practice to the Web environment and making library data available for exchange and reuse on the Semantic Web. As the library community looks outward and forward, library standards and technologies are converging with Web practices in three areas: content description, data models, and data exchange. This article captures the essence of the core standards and technologies that underlie the daily work of practitioners of library service, including Resource Description and Access, Functional Requirements for Bibliographic Records, the Linked Data environment, Resource Description Framework, and the Bibliographic Framework Transition Initiative. The article will discuss their intersections with existing practice during this period of transition as well as their potential impacts on the future cataloging practice.

KEYWORDS *cataloging, Linked Data, Resource Description and Access, Resource Description Framework, Bibliographic Initiative Framework, Semantic Web, entity-relationship models*

The library community is poised to witness a revolutionary moment for library technology. In the past half-century, the electronic library catalog evolved on a parallel path with publicly-accessible networks, such as the World Wide Web. The OPAC and discovery systems have served as distinct points of contact, but library data continue to reside largely in isolated library

Received 14 January 2013; accepted 9 April 2013.
Address correspondence to Kimmy Szeto, Newman Library, Baruch College, 151 East 25th Street, New York, NY 10010. E-mail: ks287@columbia.edu

silos. Recent developments in several library technical standards will bring about a grand technological convergence with the Web in three areas: (1) content description, (2) data models, and (3) data exchange.

These new standards will change the way resources are described as well as how library data will be structured, accessed, and used by the library community and beyond. There is a wealth of technical literature on these technical topics. (A list of further reading is provided at the end of this article.) In this article, I aim to capture the essence of these technologies that underlie the daily work of practitioners of library service and to discuss the potential impacts on cataloging practice these developments might bring. I will clarify some potentially confusing terminologies and illustrate some concepts graphically. I will also explain how some of these technologies will intersect and overlap during this period of transition.

AN OVERVIEW OF RECENT ADVANCES IN RESOURCE DESCRIPTION STANDARDS AND TECHNOLOGIES

Already upon us is the new content description standard Resource Description and Access (RDA), which will replace the Anglo-American Cataloging Rules, 2nd edition (AACR2). RDA is modeled on the Functional Requirements for Bibliographic Records (FRBR), which so far has existed as a concept with vast potential to revolutionize resource description but has encountered difficulties reconciling with AACR2 concepts, rules, and terminologies. RDA represents a revision to AACR2 that implements FRBR concepts and incorporates FRBR terminology. The convergence of RDA and FRBR will focus cataloging, or resource description, on the resources' relationships with each other and steer the process of retrieval and access toward navigating links through a hierarchy of relationships. This change in emphasis positions libraries to participate in the emerging Semantic Web.

The Semantic Web aims to generalize the World Wide Web, which mostly consists of hyperlinked documents, to a web of data, where Web resources are systematically linked so that machine processing can yield meaningful knowledge and inferences about the data for human consumption. The set of underlying principles that allow this systematic structure to exist is Linked Data. The concept is not new; it has been widely applied in areas such as controlled heading in library catalogs (since the nineteenth century!), database models, filing systems, and through the network of scholarly citation practices. RDA converges with Linked Data principles by complying with the Resource Description Framework (RDF), the data model required for Linked Data. In a parallel effort to the development of textual guidelines and instructions, RDA is also designed as a set of RDF classes and properties, as well as associated vocabularies. Expressing library data in a manner compliant with Semantic Web standards will foster information exchange and reuse in the broader web of data outside the library world.

The data model supported by RDA will also provide the basis for migrating legacy library data currently isolated in MARC-based systems. The Library of Congress (LC) launched the Bibliographic Framework Transition Initiative (BIBFRAME) to develop the successor to the current Machine-Readable Cataloging exchange format (MARC 21). Complementing RDA, BIBFRAME will represent the third area of convergence. BIBFRAME will allow bibliographic data to be encoded as linkable data, a vehicle for accessibility on the Semantic Web.

CONVERGENCES IN RESOURCE DESCRIPTION TECHNOLOGIES AND STANDARDS

Convergence 1: Modeling RDA on FRBR

FRBR

FRBR is a conceptual model developed by the International Federation of Library Associations and Institutions (IFLA). The IFLA Study Group described a generalized conceptual model that establishes entities and relationships for information objects (IFLA 1997). Until then, terms such as "book," "edition," "publication," "work," and "item" were not always precise. When we refer to "this book," are we referring to this particular object (a FRBR Item) or all the identical printed copies of this particular publication (a FRBR Manifestation)? Or are we referring more broadly to the collective conceptual contents in the book (a FRBR Work)? Or are we referring to this work in the text form, in a specific language (a FRBR Expression)? These are clearly delineated in a hierarchy in FRBR's Group 1 entities.

FRBR's Group 2 entities consist of Person and Corporate Body. Group 2 and Group 1 entities are linked through specific relationships, much like relating a title to an author, to an illustrator, to an editor, and so forth. An important attribute of each Group 2 entity is the Role. Role provides the relator term that often goes unrecorded in bibliographic records. It is a crucial piece of information that supplies meanings to relationships in the Semantic Web environment.

FRBR's Group 3 entities supply Concept, Object, Event, and Place information to any Group 1 or Group 2 entity, much like applying subject headings to a title or authority data about a person.

In addition to establishing these inherent relationships, FRBR also delineates content relationships. We encounter content relationships in reprints, photocopies, and microform (equivalent relationship); in translations, revisions, and arrangements (derivative relationship); in parodies and adaptations to a different genre (a new Work related to the original); in reviews, criticisms, and annotations (descriptive relationship to the original); as well as between issues and serials (whole/part relationship); and for issues of serials, books with CD-ROMs, and music score and parts (sequential, accompanying, and companion relationship). Catalogers often encounter these relationships and make cataloging decisions that ultimately affect users' catalog search results. For example, will a user find Shakespeare's *Romeo and Juliet* in a search for Bernstein's *West Side Story*? (For graphical illustrations of FRBR's three groups of entities and bibliographic relationships, refer to Barbara Tillett [2004].)

FRBR relates the bibliographic model with user processes through four User Tasks: Find, Identify, Select, and Obtain. These user tasks reiterate Charles Cutter's objectives of a library catalog (Cutter 1875) as users navigate through all the interrelated FRBR entities and complete FRBR's holistic approach to the bibliographic universe.

RDA Based on FRBR

The FRBR model has had great influence on the development of RDA. In fact, in the early stages of revising AACR2, one of the principles of what was then AACR3 was to incorporate FRBR terminology and concepts (Joint Steering Committee for the Development of RDA [JSC] 2005a), and to align the organization of RDA with the FRBR model (JSC 2005b; JSC 2009). However, without a viable data model, RDA would remain a content standard for producing textual statements. A Web-ready, Web-scale data model is necessary for the next convergence.

Convergence 2: RDA in RDF: Library Data as Linked Data

RDA in RDF

Another significant corollary of RDA's convergence with FRBR is moving resource description from flat sequences of statements to the application of the hierarchical, entity-relationship model. Not only does the model facilitate user tasks, it also aligns library cataloging practice to the general model of relational database design, as well as the data model underlying the Semantic Web. To facilitate this, the developers of RDA formalized RDA's element set based on the RDF data model. RDF, developed by the World Wide Web Consortium (W3C), is compliant with a host of Semantic Web technologies, such as OWL (Web Ontology Language), RDFS (RDF Schema Language), and SKOS (Simple Knowledge Organization System) vocabularies. This extensibility is central to RDA's aim of opening library data to the wider Semantic Web.

RDF works well with FRBR in that they both subscribe to the entity-relationship model when drawing relationships between resources. Additionally, it provides a data structure for meaning to be embedded and for knowledge to be built on these embedded meanings. RDF's formalism is quite complex (Klyne and Carroll 2004), but the concept is surprisingly straightforward. RDF relationships are expressed in statements that take this form:

THING → is in a RELATIONSHIP with → a different THING

This RDF statement, also known as the RDF triple, has three parts: two "things" and one unidirectional relationship in between (all three parts carry equal weight). This concept is widely used in fields like mathematics, linguistics, and database design, so the nomenclature varies for the RDF triple. Here are some common terms that are roughly equivalent:

Database: Entity → Relationship → Entity
Sentence: Subject → Predicate → Object
Directed Graph: Node → Arc/Edge → Node
Cataloging: Resource → has Property → Value

In Web implementations, the most commonly used method to express these relationships is Extensible Markup Language (XML), but there are other ways to visualize RDF. It can be shown using mathematical expressions in tabular form, or as a labeled, directed graph. Since RDF is conceptual, visualizations cannot always fully represent the concepts and can pose limitations legibility. XML and mathematical expression are difficult for the human eye to read (Figures 1 and 2); the tabular form is limited to showing relationships in two dimensions (Figure 3). The labeled, directed graph offers the most flexibility for the purpose of illustration (Figure 4).

Linked Data Implementation

Linked Data

To implement the RDF model on the Web and link all these entities and relationships together over the Web, we must add a few standards and

```
{Work, titleOfTheWork, Hamlet}
{Work, Author, Person}
{Person, nameOfThePerson, "Shakespeare, William, 1564-1616"}
```

FIGURE 1 RDF triples as mathematical expressions.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!--This fragment is created for illustration purposes only and is not intended to
demonstrate any encoding standard.-->
<rdaSimple
       xmlns:rda="http://rdvocab.info/Elements/"
       xmlns:rdg2="http://rdvocab.info/ElementsGr2/"
       xmlns:rdg3="http://rdvocab.info/ElementsGr3/"
       xmlns:frbr="http://rdvocab.info/uri/schema/FRBRentitiesRDA"
       xmlns:role="http://rdvocab.info/roles/"
       >
   <frbr:Work uri="http://rdvocab.info/uri/schema/FRBRentitiesRDA/Work">
      <rda:titleOfTheWork>
           Hamlet
       </rda:titleOfTheWork uri="http://rdvocab.info/Elements/titleOfTheWork">
       <role:author uri="http://rdvocab.info/roles/authorWork">
         <frbr:Person uri="http://rdvocab.info/uri/schema/FRBRentitiesRDA/Person">
           <rdg2:nameOfThePerson uri="http://rdvocab.info/ElementsGr2/nameOfThePerson">
             <frbr:Name uri="http://rdvocab.info/uri/schema/FRBRentitiesRDA/Name"
             id="http://id.loc.gov/authorities/names/n78095332" vocab="LCNAF">
               Shakespeare, William, 1564-1616
             </frbr:Name>
           </rdg2:nameOfThePerson>
         </frbr:Person>
       </role:author>
   <frbr:Work>
</rdaSimple>
```


FIGURE 2 RDF triples expressed in XML.

Title	Author
Hamlet	William Shakespeare
Hamlet	Peter Ilich Tchaikovsky

Name	Date of Birth	Authoritative Label
William Shakespeare	1564	Shakespeare, William, 1564-1616
Peter Ilich Tchaikovsky	1840	Tchaikovsky, Peter Ilich, 1840-1893

FIGURE 3 RDF triples shown as tables.

ways of identifying and addressing these objects. This is where Linked Data Principles come in.

The idea of linking one set of information to another through an identifier, a number, or a keyword is well established. Programmers and developers use pointers in programming languages and draw relationships in relational databases. In everyday use, whether in the electronic or analog environment, arranging and filing documents and records invariably includes elements of linking data. In business, the customer number on an invoice links to the customer's account; manufacturer numbers link to product descriptions and inventory. In libraries, we link bibliographic records to holding and item records, and we link names and subjects to authority files.

On the World Wide Web, the concept of linking has most often applied to hyperlinking between documents (i.e., linking between discrete files

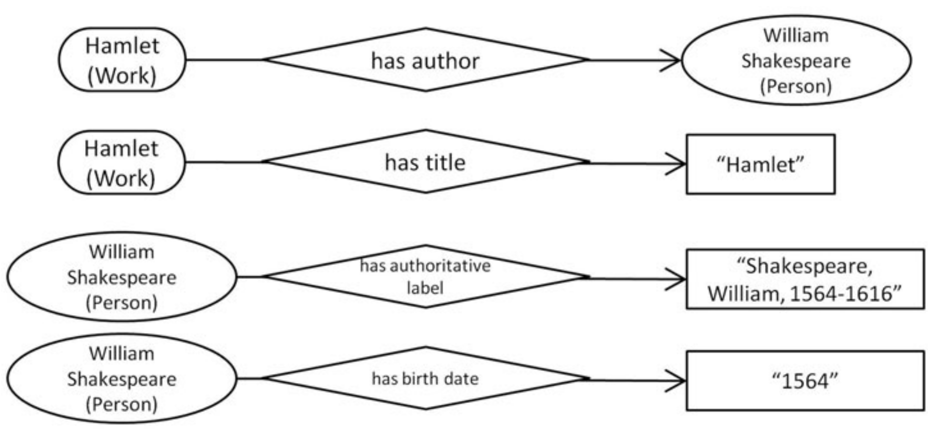


FIGURE 4 RDF triples shown as labeled, directed graphs. Rounded shapes represent entity references; diamonds represent relationships; rectangles represent entity literals.

containing text or binary data using a URL). This use of hyperlinking shows that a relationship exists between two documents but does not specify the nature of the relationship. In the generalized Linked Data environment, relationships carry meaning (or semantics), and objects are not limited to documents but can refer to any data, any "thing"—entities like concept, place, event, and person, and relationships themselves are objects as well. The resulting Web will then be more "meaningful," or semantically rich. But of course, content providers and users of this Semantic Web must follow a set of common principles for the design, structure, and presentation of data, otherwise the data will simply be ignored. Tim Berners-Lee was first to articulate the four Linked Data principles as a set of best practices for the Semantic Web (Berners-Lee 2006):

1. Use URIs as names for things.
   The unique identifier is a familiar cataloging concept. It is also fundamental to Semantic Web technology, where every resource must be named with a globally unique Universal Resource Identifier (URI) in order to distinguish it from any other resource.
2. Use HTTP URIs so that people can look up those names.
   These URIs are required to be in a specific form according to the HyperText Transfer Protocol (HTTP). URLs, widely used on the Web for locating documents, are one type of HTTP URI.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
   Berners-Lee (2006) recommends HTTP URIs so that the identifier can also serve as the Web locator that allows users to retrieve the definition and other related information about a resource, a feature known as dereferencing. Standards such as RDF and the SPARQL Protocol and RDF Query Language (SPARQL) allow semantics to be embedded in the data. RDA, in particular, works with RDF.
4. Include links to other URIs so that they can discover more things.
   When a critical mass of online data sets follows these principles, the Linked Data environment will enable large-scale integration and reasoning over globally connected data sets on the Semantic Web.

Expressing RDF as Linked Data

When Linked Data principles are applied to the RDF model, each of the three parts of an RDF triple must have a HTTP URI. But this is not always possible in implementation. Therefore, the formulation of RDF entities allows for three forms: reference (using URI), literal (text), and blank (placeholder). The "blank" accounts for the situation where an entity is created but no name is assigned (e.g., as an intermediate step in a metadata creation workflow or when a resource is a collection of further links but itself contains no useful information). Some information must be exact text, such as transcribed statements, and this text is represented as literals. Some literals such as date, LCCN, ISBN, and so forth can be consistently formatted for machine processing.

While optimal for machine processing, URIs' long strings of text are not useful for human use. For convenience of human manipulation, URIs may have lexical labels. These lexical labels are, in fact, what we have been using for authorized forms. But unlike URIs, labels are not inherently unique. On the other hand, URIs are required to be unique. Therefore, in the Linked Data environment, URIs replace lexical labels as authorities.

EXPRESSING RDA IN RDF

Figures 3 and 4 illustrate an example of how to interpret information in tabular form and re-imagine it as a labeled, directed graph. In Figure 3, the table contains a list of information about various books. Before we look at specific data entries, consider that this entire table represents a group of similar entities. These similar entities are grouped together into Classes, and there is an entity class we now call Works. Within the table, each column header represents one particular property (e.g., the Author column in Figure 3 is represented as the "has Author" property in Figure 4). These authors can be grouped together as the entity class Person. As we fill in actual data for each of the rows, the resulting triples are called Instances of the triple.

LINKING LIBRARY DATA

To further illustrate Linked Data principles at work, examine Figure 4. Notice that there are several entities that are repeated. These entities, which are identical, can be combined to form a single, larger graph (Figure 5). As more information becomes available, entities that are identified as identical or equivalent can be merged in the same way into the graph, and additional links to other resources will be added to the graph. As a result, the graph will expand and can expand infinitely.

In this environment, bibliographic description is no longer based on the bibliographic record as a unit. The unbounded nature and the ability to merge are the key features of the Linked Data environment that make it possible for the library catalog to interface with data outside the library community and become a part of the global Web of knowledge. Different data sources that use equivalent concepts and relationships can be seamlessly integrated by direct human intervention or by computer programs; resources and concepts that are not exactly identical can be mapped algorithmically by reasoning engines.

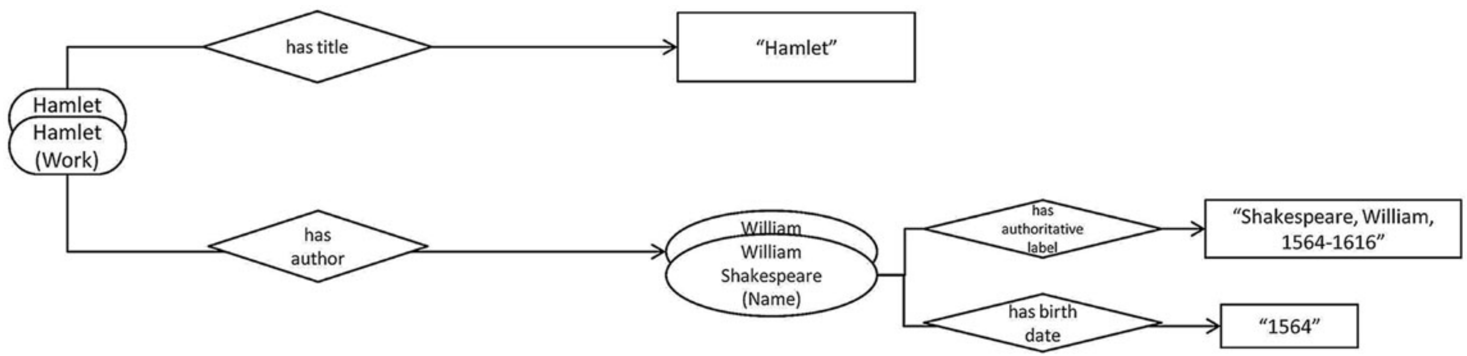


FIGURE 5 A single graph combining all the triples in Figure 4.

Cataloging as Linked Data

To implement a cataloging practice that reflects RDF triples as in Figure 5, all the resources and properties must have a unique URI. But what are these URIs? What are the definition and scope of these resources? What exactly is a Work? What exactly is a Person? Where are these terms dereferenced? More broadly, can existing cataloging standards be carried over into this new data model?

Figure 6 shows the underlying URIs referencing RDA properties and LC authorities. RDA's RDF properties and entity classes are formally defined and published in the Open Metadata Registry ("RDA Vocabularies" 2013; Hillman 2010), and LC authorities and vocabularies are published by the Linked Data Service (http://id.loc.gov). These registered entities define how these properties are to be used and how they are related to each other internally and to external entities, such as authority data, controlled vocabularies, thesauruses, or code lists. The metadata registries will guide catalogers in their formation of graphs as well as provide the necessary URIs, and ultimately make RDA properties available for other communities to use for such purposes as real time validation, synchronization, and application development (Bizer, Heath, and Berners-Lee 2009). This puts the library community in a position to enter the broader Semantic Web.

Convergence 3: Library Data Exchange Format for the Semantic Web

MARC for Linked Data?

Most elements of the MARC format have been analyzed and mapped to RDF properties (Coyle 2011; Hillmann and Dunsire 2012). However, because MARC was developed over four decades ago to automate the creation and printing of catalog cards, the format presents many structural limitations in adapting to the Web-based environment (McCallum 2012). RDA, unlike AACR, is not limited to standardizing description for the purpose of printed catalog cards. Emphasizing the recording of bibliographic data over composing textual records, RDA is a general-purpose content standard with the flexibility to lend itself to any form of data presentation and transmission. So, while RDA is fully supported in MARC 21, it is possible to pair it with other exchange formats. This flexibility encourages the library community to reimagine a bibliographic environment that fuses seamlessly with the Linked Data environment.

BIBFRAME

BIBFRAME, the next data exchange format to replace MARC, is still in development at the conceptual level. Like RDA, it will be developed as an

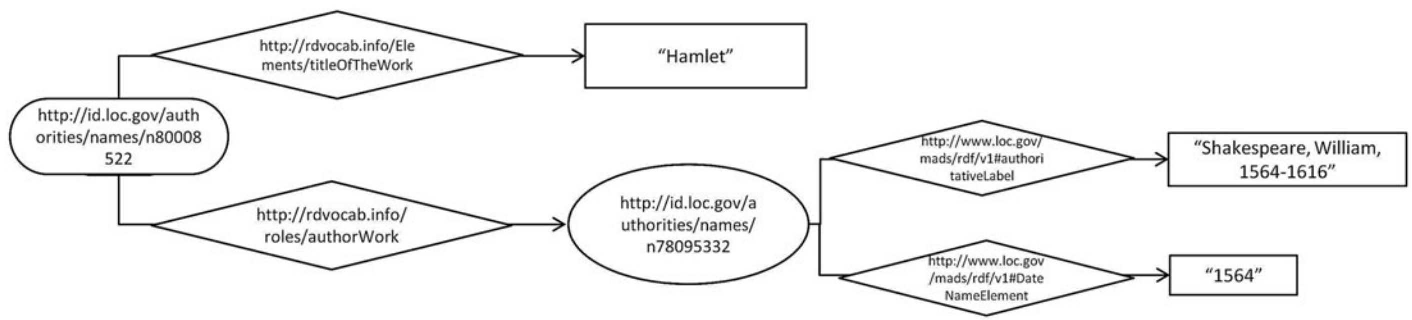


FIGURE 6 The graph in Figure 5, but showing URIs instead of lexical labels.

independent model. BIBFRAME will definitely work with RDA, but it can also work with other content standards like DACS (Describing Archives: A Content Standard) and CCO (Cataloging Cultural Objects). BIBFRAME will continue to support bibliographic description, authority data, holdings, and classification, and will be able to address all types of holdings, including digital and born-digital materials. In the meantime, BIBFRAME will take advantage of the groundwork that has been laid on the content standards and linked data fronts, such as making use of URIs wherever applicable.

Data in BIBFRAME fall into four core classes: Work, Instance, Authority, and Annotation. (To set BIBFRAME concepts apart from FRBR concepts, the prefixes BF and FRBR will be used.) According to a preliminary report (LC 2012), BF Work encompasses "the conceptual essence of the cataloged item"; BF Instance reflects "an individual, material embodiment of the [BF] Work." Note that BIBFRAME's Work-Instance structure is slightly different from that of FRBR: BF Work encompasses both FRBR Work and FRBR Expression, whereas BF Instances are very similar to FRBR Manifestation. BF Authority associates "key authority concepts that have defined relationships reflected in the [BF] Work and [BF] Instance." Examples include topics, people, institutions, and places. BF Annotation is a resource that "decorates other BIBFRAME resources with additional information." This is a new concept for additional assertions, such as reviews, abstract, and excerpts (relating to BF Work); holdings, book cover images, and tables of contents (relating to BF Instance); and authority information and administrative metadata (relating to BF Annotation) to be actively sought out for library data. (For graphical illustrations of BIBFRAME's entities, refer to the preliminary report [LC 2012].)

The notable difference of BF Annotations from the other core classes is that it is not designed to be controlled, but is either created locally or drawn from the Web. The notion of BF Annotation will decentralize data, augmenting bibliographic data with selected external sources as well as local resources and user-generated tags or comments. Current cataloging practice already allows a limited amount of external data (e.g., external links in the MARC 856 field). However, including external data as a central component has the potential to redefine the scope and goals of cataloging. Figure 7 shows the composition of a future "catalog record" that will be searched and displayed. In addition to providing resource description according to content description rules, a number of annotations will be gathered from around the Web.

Iterative development of the BIBFRAME model is ongoing; testing of mappings and conversion tools is in the preliminary stages. Even though BIBFRAME is designed to be format-neutral, its development will focus on ensuring compatibility with RDA and RDF.

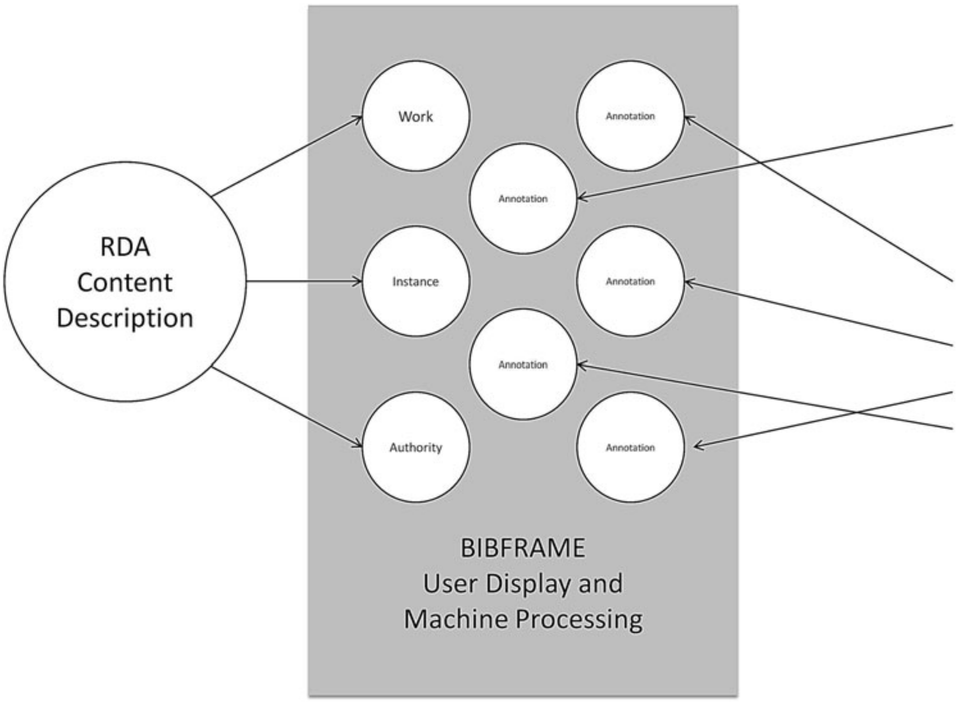


FIGURE 7 The future "catalog record" (in gray), according to the BIBFRAME structure, will include content description as well as a number of annotations gathered from around the Web.

THE FUTURE OF RESOURCE DESCRIPTION

These technological convergences all aim to expose library data to the wider web of data for exchange and reuse, as well as to enrich library data with resources from the Semantic Web. What will the work of a cataloger be like in the future?

For now, catalogs will continue to use MARC and to type lexical labels in textual descriptions. As RDA adoption becomes more widespread, we will begin to adapt to a new way of conceiving of resource description. RDA has laid the groundwork; authority data and vocabularies are ready to be employed.

Resource description of the future will be done on the Web, for the Web. We will no longer be concerned with whether the entry will print, nor will we select the main entry over added entries. With the expanded availability of authorities, we will spend more time looking for suitable vocabularies and linking to them. The shift of emphasis on recording the URI over lexical labels might mean less typing but more drag-and-drop and prune-and-graft, should vendors develop the graphical interface. Catalogers' role will be more akin to that of data curators. They will establish search policies for external resources from the entire Web, configure systems to harvest these external data, and set the scope of the local display.

REFERENCES

Berners-Lee, Tim. 2006. "Linked Data Design Issues." World Wide Web Consortium. Last modified June 18, 2009. http://www.w3.org/DesignIssues/LinkedData.html.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2009. "Linked Data: The Story So Far." *International Journal on Semantic Web and Information Systems* 5 (3):1–22.

Coyle, Karen. 2011. "MARC21 as Data: A Start." *Code4Lib* 14. http://journal.code4lib.org/articles/5468.

Cutter, Charles A. 1875. *Rules for a Printed Dictionary Catalogue*. Washington, DC: United States Government Printing Office.

Hillmann, Diane. 2010. "RDA Vocabularies: Process, Outcome, Use." *D-Lib Magazine* 16:1/2. http://dlib.org/dlib/january10/hillmann/01hillmann.html.

Hillmann, Diane, and Gordon Dunsire. 2012. "MARC21 as RDF." Presentation at the MARC Format Interest Group Meeting at the American Library Association Midwinter Meetings, Dallas, TX, January 21. http://www.slideshare.net/smartbroad/mfig-on-mar21rdf.

IFLA Study Group on the Functional Requirements for Bibliographic Records. September 1997, amended and corrected through February 2009. *Functional Requirements for Bibliographic Records: Final Report*. The Hague: International Federation of Library Associations and Institutions. http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

Joint Steering Committee for the Development of RDA. 2005a. "Draft of AACR3 Part I: Background to the December 2004 Draft." January 12. http://www.rda-jsc.org/aacr3draftpt1.html.

———. 2005b. "Outcomes of the Meeting of the Joint Steering Committee Held in Chicago, USA, 24–28 April 2005." May 19. http://www.rda-jsc.org/0504out.html.

———. 2009. "RDA: Resource Description and Access: Scope and Principles." July. http://www.rda-jsc.org/rda.html#scope.

Klyne, Graham, and Jeremy J. Carroll, eds. 2004. "Resource Description Framework (RDF): Concepts and Abstract Syntax." World Wide Web Consortium. http://www.w3.org/TR/rdf-concepts/.

Library of Congress. 2012. "Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services." Library of Congress. November 21. http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf.

McCallum, Sally. 2012. "Bibliographic Framework Initiative Approach for MARC Data as Linked Data." Presentation at IGeLU, Zürich, Switzerland, September 13. http://www.loc.gov/marc/transition/pdf/BFI-IGeLU-ppt2003.pdf.

"The RDA (Resource Description and Access) Vocabularies." 2013. *Open Metadata Registry*. Accessed April 1, 2013. http://rdvocab.info.

Tillett, Barbara. 2004. *What is FRBR? A Conceptual Model for the Bibliographic Universe*. Washington, DC: Library of Congress, Cataloging Distribution Service. http://www.loc.gov/cds/downloads/FRBR.PDF.

FURTHER READING

RDA

Coyle, Karen. 2010. "RDA Vocabularies for a Twenty-First-Century Data Environment." *Library Technology Reports* 46 (2). Chicago: ALA TechSource.

Hillmann, Diane, Karen Coyle, Jon Phipps, and Gordon Dunsire. 2010. "RDA Vocabularies: Process, Outcome, Use." *D-Lib Magazine* 16: 1/2. http://dlib.org/dlib/january10/hillmann/01hillmann.html.

Joint Steering Committee for the Development of RDA. 2013. Last modified January 18. http://www.rda-jsc.org/aacr3draftpt1.html.

FRBR

Carlyle, Allyson. 2005. "FRBR: Challenges for Implementation in AACR2, with Some Attention to Nonbook Mmaterials." Presented at FRBR in 21st Century Catalogues: An Invitational Workshop, Dublin, OH, May 2–4, 2005. http://www.oclc.org/content/dam/research/activities/frbr/frbr-workshop/presentations/Carlyle_IFLA_FRBR_May_2005.ppt.

Carlyle, Allyson. 2006. "Understanding FRBR as a Conceptual Model: FRBR and the Bibliographic Universe." *Library Resources & Technical Services* 50 (4):264–73.

Danskin, Alan, and Ann Chapman. 2003. "Bibliographic Records in the Computer Age." *Library & Information Update* 2 (9):42–3.

IFLA Study Group on the Functional Requirements for Bibliographic Records. September 1997, amended and corrected through February 2009. *Functional Requirements for Bibliographic Records: Final Report*. The Hague: International Federation of Library Associations and Institutions. http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

Madison, Olivia. 2005. "The Origins of the IFLA Study on Functional Requirements for Bibliographic Records." In *Functional Requirements for Bibliographic Records (FRBR): Hype, or Cure-All?*, edited by Patrick Le Boeuf, 15–37. Binghamton, NY: Haworth Press.

Taylor, Arlene, ed. 2007. *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*. Westport, CT: Libraries Unlimited.

Tillett, Barbara. 2004. *What Is FRBR? A Conceptual Model for the Bibliographic Universe*. Washington, DC: Library of Congress, Cataloging Distribution Service. http://www.loc.gov/cds/downloads/FRBR.PDF.

Paper 3

The mystery of the Schubert song: the linked data promise

This publication first appeared in
Notes, the Quarterly Journal of the Music Library Association, 74(1), pp. 9-23.

DOI: 10.1353/not.2017.0071

THE MYSTERY OF THE SCHUBERT SONG: THE LINKED DATA PROMISE

By Kimmy Szeto

MYSTERY OF THE SCHUBERT SONG

"I think the German group really needs one more song." A music reference situation unfolded as the voice teacher discussed a recital program with her student. "I remember that Austrian soprano. . . . Was she Austrian? What's her name? Strada? Estrada? The last song on her album is a Schubert song. . . . It's upbeat; it starts on a high G. What's that song called? It's one word. . . . I think it ends with '-*lein*'. . . ." The teacher thumbed through her volumes of the complete Schubert songs, then started running her finger down the index. In the meantime, the student picked up her mobile phone and pulled up "List of songs by Franz Schubert" on Wikipedia. She moved on when she realized the songs were listed by opus and catalog numbers. Then she pulled up "List of compositions by Franz Schubert by genre," and scrolled to the section "Lieder with piano accompaniment," first the eleven cycles and sets, and then three dozen or so by voice type. At that point, she looked overwhelmed by the sight of the remaining list of 500 or so entries.

I stayed silent as the accompanist should, estimating a maximum of thirty seconds before they would both give up. But my librarian persona leapt into action. I pulled out my mobile phone, searched for "soprano obituary strada" (without the quotes). I realized the name was misspelled (thanks to Google's "Did you mean…" suggestion).[1] The actual name was Stader and she was Swiss, not Austrian. "Maria Stader?" I asked. "Right, Maria," the teacher said, still scanning the index. I then searched for "maria stader schubert LP." Among the top results were two entries on Discogs.com, an online marketplace for music collectors. "Was it a

Kimmy Szeto is an assistant professor and metadata librarian at Baruch College, City University of New York, where he oversees metadata management for digital resources. His recent research focuses on the technical and conceptual tensions between cataloging practice and the linked data environment. He is an active presenter at Music Library Association meetings, and has published in the *Journal of Web Librarianship*, the *Journal of Electronic Resources Librarianship*, and the *Encyclopedia of Information Science and Technology*. Outside the library and academia, Kimmy can be heard as the chamber arranger of symphonic works and as a collaborative pianist in theaters around New York City.

URLs cited herein accessed 18 May 2017

1. Google's "Did you mean…" feature uses multiple probabilistic and machine-learning algorithms that are based, in part, on the user's search history. So, this search is not meant to be replicable. It was with an element of chance that I hit upon a good suggestion.

live recording of a concert?" The teacher did not think so. So I chose the Discogs entry for the 1958 Deutsche Grammophon studio album *Liederabend*.[2] The only song title that ended with "*-lein*" was not the last song on either side of the LP. It had more than one word, and it was not even by Schubert (*Des bescheidene Wünschlein* by Othmar Schoeck). But I spotted another track and took a leap of faith: *Seligkeit*?

"How did you know?"

I did not. But I knew how to look, I knew when to ask follow-up questions, and I knew when to guess. While retrieving *Seligkeit* on IMSLP.org, I remarked that to train as a singer nowadays was to train as a librarian. It turned out the high G-sharp was not at the start of the song, but it was the start of the last phrase of the song. Nonetheless, the song was exactly the one the teacher was looking for, and it was perfect for her student's recital program.

CATALOG SEARCHING AND DATA CONNECTEDNESS

While singers might benefit from information literacy skills, they should not need librarianship training. But in so many situations like this one, absent a reference librarian, our online services fall short.

Our current bibliographic systems can respond to a search for an LP as the material format, Schubert as the author, and solo songs with piano as the subject.[3] This search yields over one thousand results. Adding "strada" or "estrada" to the search, however, yields zero result.

At this point, an experienced searcher would focus on revising the singer's name. Remarkably, adding the correct name "stader" to the search reduces the results to five, and includes the appropriate answer. The problem is getting to the correct name. Neither the WorldCat public interface nor the Library of Congress offers a name search by nationality. One could attempt to add more keywords: adding "soprano" would yield over four hundred results, but adding "Austria" or even the correct country "Switzerland" would not yield bibliographic records of any of Stader's recordings, because such a search would look for keywords in bibliographic records, and nationalities are recorded in a separate authority file.

A persistent searcher might at this point use the fact that the recording was old, and comb through all the results in chronological order. Going down this path would require examining a minimum of seventy-two bibliographic records before reaching the one for an album titled *A Maria*

2. Maria Stader, Franz Schubert, Felix Mendelssohn-Bartholdy, Othmar Schoeck — *Liederabend*, Deutsche Grammophon LPEM 19136 (1958), LP, https://www.discogs.com/release/7821855.

3. A WorldCat search with the command "mt=lps au:schubert su:songs with piano"; in the rest of this article, the commands "pn: " and "kw: " are used for WorldCat searches for names and keywords, respectively.

Stader Recital, which includes the same tracks as *Liederabend*.[4] Even then, there is no guarantee the searcher would recognize the spelling discrepancy in the singer's name to select the record for further evaluation.

Choosing another path, a savvy searcher might hone in on the name and nationality of the singer using the VIAF: Virtual International Authority File,[5] but would still come up empty: a search for "soprano" yields too many results, but narrowing down by "Austrian," "Strada," or "Estrada" does not provide any further clues. The VIAF record for Maria Stader does turn up if the search phrase includes "Swiss." In other words, this is a dead end where searching with the wrong nationality could not correct the name, and searching with the wrong name could not correct the nationality. The searcher is eventually turned back to square one: namely, examining an index, be it the chronological list of sound recordings, the list in a Wikipedia article, or the title index in the Schubert *Complete Songs*.

Can bibliographic systems do better than this? One possible technological solution is linked open data.

SEARCH STRATEGIES AND DATA STRATEGIES

Linked open data is a set of design principles for making data freely available on the Internet in a structure that allows machine processing to understand, connect, and enrich the content represented in the data.[6] This web of machine-parsable data enables the creation of new knowledge as machines make inferences based on integrating existing data sets from disparate sources.[7] Could linked open data enable machines to solve the mystery of the Schubert song? Very likely, had data from WorldCat, VIAF, and Discogs been available as linked data for machines to make inferences beyond the known, and somewhat incorrect, information.

In the search for *Seligkeit*, the teacher and the student both tried to browse a title index, based on two pieces of data (Schubert and song). Had they been in a library, a reference librarian might, at first, try the catalog searches discussed above, based on several more pieces of data (Schubert, song, singer's name, singer's nationality, format of recording, date of recording). This is not to say that the index browse and the catalog search

4. *A Maria Stader Recital*, Decca DL 9994 (1958), OCLC WorldCat no. 2764096. The WorldCat record no. 30023258 for the *Liederabend* album appears eighteen records later.

5. http://www.viaf.org. Even though Maria Stader's nationality is recorded in the Library of Congress Name Authority File, searching the public interface on id.loc.gov yields no results. To perform a search for the field that includes nationality requires a tool such as the Connexion software search with an OCLC authorization credential.

6. I use the term "machines" to refer to computers, as well as all other computing devices, learning machines, and neural networks.

7. "Data" refers here to individual pieces of data, and "data set" refers to pieces of data grouped together into a machine-readable structure.

could not have led to the answer. But six hundred songs are not easily browsed, and the catalog search was able to reduce that figure by only a fraction. My strategy, given only a mobile phone and thirty seconds, involved looking, in a particular order, for three pieces of data: Who was this singer (name)? Which LP recording was it (Schubert)? Which track in the recording was it (title ending with "-*lein*")? The reason for this particular order was to narrow down answers as quickly as possible, so that I could take a guess before time was up. This two-searches-and-a-guess strategy was neither unusual nor unique, but could machines have come up with it? How can we make more use of machines as an analytical tool? Machines are only as good as the programs we run and the data we supply, and there is much the library community can do about the data. After all, creating and managing data is one of our areas of expertise.

Computers are machines designed to perform arithmetic and logical instructions on data. Through a process called decomposition, humans translate complex problems into sequences of simple machine instructions, and break down data into machine-parsable sets. The simple and repetitive nature of computing works well when we supply data sets with a uniform structure in which what the data represent is unambiguous and atomic, that is, already in the lowest level of detail. In our search for *Seligkeit*, the voice teacher offered several pieces of information: a soprano (with a possible name and a possible nationality), an LP with a Schubert song (with a possible portion of the title). Figure 1 shows one possible way to decompose the data based on my search strategy.[8]

The problem, and the challenge, is to start with these pieces of data and somehow end up with the song title *Seligkeit*. Web searching, for the most part, means to enter the data as text strings, and look for where they appear on Web pages. Catalog searching finds records that have these text strings in particular fields. While field searching in a library catalog is more precise, the results are limited to bibliographic records in library systems. Linked open data, on the other hand, not only offer a global web of data for field searching, but also allow computer programs to evaluate and return additional data that ordinarily would fall outside the scope of Web and catalog searching.

DATA LINKING ON THE INTERNET

The World Wide Web connects hypertext documents via hyperlinks, and has grown from a handful of pages when first implemented in 1991

8. Although they are legitimate clues, I did not consider the high G or the upbeat nature of the song, because I knew that this information was unlikely to turn up in a Web-based search or be recorded in library data.

a person			
	who performs as		
		a soprano	
	who resides in		
		Austria	*or possibly somewhere in Europe*
	who has the name		
		Strada	
		Estrada	*or possibly something with a similar sound or spelling*
a song			
	having been composed by		
		Franz Schubert	
	has instrumentation/voices consisting of		
		voice (solo)	
		piano	
	has the title		
		that has one word	*possibly plus an initial article*
			possibly with more words it's most likely not a long title
			possibly has the word that ends with "-lein*"*
			or something similar to that
an LP			
	was recorded on a date		
		not too recent	
	includes a track that is		*possibly on the last track on one of the sides*
		the song above	
		performed by	
		the person above	

Fig. 1. A decomposition of data involved in the mystery of the Schubert song (data in bold are without uncertainty)

to over one billion Web sites thirty-six years later.[9] Now imagine a similar scale of connected data sets! The concept behind achieving a vast amount of data interconnectedness is surprisingly simple. Basically, it requires a critical mass of data sets to appear on the Internet following four design principles. They are listed in figure 2,[10] along with current technologies[11] that satisfy their purposes. Their ramifications are elaborated below.

9. A Web-site counter with references to the counting algorithm can be found on "Total Number of Websites," *Internet Live Stats*, http://www.internetlivestats.com/total-number-of-websites.

10. Table adapted from Tim Berners-Lee, "Linked Data," *Design Issues*, last modified 18 June 2009, https://www.w3.org/DesignIssues/LinkedData.html. Tim Berners-Lee describes these four characteristics as "expectations of behavior" that are often erroneously understood as rules or requirements. URI, HTTP, RDF and SPARQL are listed not as requirements but as technologies of choice for their already widespread use on the Internet. He explains these brief design notes more fully in his presentation "Tim Berners-Lee: The Next Web," *TED2009*, February 2009, https://www.ted.com/talks/tim_berners_lee_on_the_next_web.

11. In this paper, "technology" refers to any application of science for practical purposes, which include computing hardware, software, as well as standards and specifications for communication protocols, data models, markup and query languages, etc.

Design Principle	Purpose	Current Technologies
Identifier	Identifier allow data and links to be uniquely identifiable, globally.	URI
Dereferencing	Dereferencing a URI is retrieving a representation of that resource. A global addressing system enables URIs to be accessed and to self-identify.	HTTP, URI
Structure and Method	Data can be useful only if queries return data. A common method or language for accessing data in a common structure makes the data globally discoverable.	RDF, SPARQL
Participation	The success of this vision of the linked data environment rests on connecting a vast amount of data across the Internet.	Include links to other URIs

Fig. 2. Design principles for linked open data

Today, we have already seen versions of these design principles in practice. The World Wide Web is a familiar example. Documents on the Web use the Universal Resource Locator (URL) as identifier; they are addressed by the prefix http:// (Hypertext Transfer Protocol); and they are marked up in a structured language HTML (Hypertext Markup Language), which provides a method (the <a> tag with the "href" attribute) to link to another document. While documents are linked on the Web, the use of the URL and HTML constrains machines from taking advantage of the ability to make inferences across data sets. As the identifier, each URL refers to the entire document, but not any content within it. Support for encoding machine-parsable data is also limited in HTML.[12] In other words, data that reside within a Web page are not well identified as data. As a result, a typical Web search is actually looking up an enormous index of text that appears on pages on the Web.

The web of data, on the other hand, will enable machines to understand what the data are about, so that, rather than just looking through indexes, machines will be able to perform reasoning and analysis.[13] The full potential of linked open data, therefore, depends on the way we make data available, or the way that the data can be identified and connected with other data via discoverable links that express an array of meaningful relationships.

12. Some metadata about the document itself can be recorded in the document header; new tags and attributes have appeared in HTML5, the latest revision of the language, which added the ability to embed custom data and designating meaning for certain types of text. But the specification document acknowledges the issue of machine processing is not adequately addressed by the language. See World Wide Web Consortium, "HTML5: A Vocabulary and Associated APIs for HTML and XHTML," last modified 28 October 2014, https://www.w3.org/TR/html5/introduction.html#introduction.

13. For example, when provided with the statements: "A soprano is a singer" and "Singers are people," the machine will be able to draw the conclusion: "A soprano is a person." Taking this example a step further, given data on names and ages of sopranos, and, from a separate data set, the gender of the names, the machine will, without explicit human input, be able to generate additional understanding, such as, "A soprano is a female person—typically; a soprano is a young male person—seldom."

In the simplest terms, providing a link between data is doing exactly that: when constructing a data set, arrange the data so that each piece of data can be connected by a link to another piece of data. The structure and method that have emerged for this purpose are the data model Resource Description Framework (RDF), and its companion query language Protocol and RDF Query Language (SPARQL). While the model is simple, the actual technical specifications are more involved, and the Web community has been developing and maintaining standards and documentation.[14]

In recent years, using RDF for constructing data sets has gained substantial traction in the library community.[15] We will delve into the details of the model after a short background discussion on this technology and its relationship with library practice.

LINKED DATA: TECHNOLOGY VS. PHILOSOPHY

Even though the design principles—identifier, dereferencing, data structure, and query language—are essential, this particular combination of technologies—URI, HTTP, RDF, SPARQL—is not required for building a web of data. Just as HTTP and HTML are not required to build a web of documents, other parallel "webs" based on other technologies exist today.[16] Essentially, the Internet provides the undergirding for multiple network technologies. No matter which "web," any Internet transmission, from the file to the software, through the computer's network cable to the modem into the Internet, triggers a cascade of interconnected and interlocking technologies that share interoperable specifications in spite of different computers, operating systems, or software applications.[17]

---

14. The suite of RDF standards is one of the many Web standards being developed and maintained by the international membership body World Wide Web Consortium (W3C).

15. An extensive report on the adoption of linked data by the library community can be found in Erik T. Mitchell, "Library Linked Data: Research and Adoption," *Library Technology Reports* 49, no. 5 (2013); as well as in his "Library Linked Data: Early Activity and Development," *Library Technology Reports* 52, no. 1 (2016). For an example of RDF use in a library linked data project, see the Linked Jazz Project (http://linkedjazz.org), developed at the Pratt Institute School of Library Information Science. A fuller technical exposition of the linked open data set built for this project can be found in Cristina Pattuelli, Alexandra Provo, and Hilary Thorsen, "Ontology Building for Linked Open Data: A Pragmatic Perspective," *Journal of Library Metadata* 15 (2015): 265–94.

16. For example, today over 140 servers with nearly five million files have been connected in "Gopherspace," a linked data environment of computer files communicated over the Internet since 1991 via the Gopher protocol and a text menu structure. (The current size of Gopherspace can be found in real time by making a query in the Gopher search engine Veronica-2: http://gopher.floodgap.com/gopher/gw?gopher/0/v2/vstat.) Another linked data environment that has been in service on the Internet since the 1980s runs on Z39.50, a communication protocol that is heavily used in the library community for its ability to perform complex, structured searches simultaneously on multiple systems. (The Library of Congress maintains the Z39.50 standard, as well as the "Z39.50 Register of Implementors," last modified September 2016, https://www.loc.gov/z3950/agency/register/entries.html.)

17. Using the Internet requires adhering to standards involving a broad range of transmission protocols, data formats, markup languages, and query languages, as well as hardware, including modems, switches, routers, and data cables.

Because the size and reach of the Internet provides a positive feedback, new Internet-related technologies, products, and services will be developed to be compatible. Initially, the popularity of HTTP and HTML made them the de facto standards for the Web. Then Cascading Style Sheets (CSS) became a ubiquitous language for Web page design and layout when major Web browsers began to support it. Similarly, in the near future, we expect technology standards for the web of data to develop and coalesce,[18] with URI, HTTP, RDF and SPARQL as the basis for this new web architecture.

By employing the Internet, we also subscribe to the philosophy behind Internet architecture that is open, interoperable, evolvable, and network-accessible. MARC, an architecture of library systems and operations since the 1970s, is at odds with this philosophy. As the Internet grew and matured, the library community long recognized the divergence between MARC—the closed architecture of library catalogs—and the open architecture of the Internet.[19] Even though MARC stands for MAchine Readable Cataloging, the central purpose of machine processing was to print database records on catalog cards and on computer screens. Since then, we continued to design databases, interfaces, and discovery systems modeled on the catalog card, and contents follow a highly controlled syntax in individually demarcated records. This design allows the library community to create quality-controlled data in robust systems that communicate with each other, but not with the open Internet. By contrast, linked data design is open and dynamic: there are no fixed records, and, at any time, any Internet user, human or machine, can supply data and create links between data. Linked data is as much a state of mind as it is technology.[20]

RECOGNIZING RDF LINKED DATA

Because of linked data's open design, a flexible data model such as RDF has emerged as the standard for the web of data. The basic structure of RDF is the triple. The RDF triple enables assertions by linking two pieces of data with a one-way relationship between the two. This

18. See Tim Berners-Lee, "Web Architecture from 50,000 Feet, Design Issues," last modified 27 August 2009, https://www.w3.org/DesignIssues/Architecture.html.

19. With over eleven thousand data elements, MARC is a closed data format that, to make it interoperate on the Internet, requires complex procedural workarounds. For an experimental study on MARC authority data, see Ionnas Papadakis, Konstantinos Kyprianos, and Michalis Stefanidakis, "Linked Data URIs and Libraries: The Story So Far," *D-Lib Magazine* 21, no. 5/6 (May/June 2015), http://www.dlib.org/dlib/may15/papadakis/05papadakis.html.

20. For an in-depth discussion on conceptualizing library data models, see Getaneh Alemu, Brett Stevens, Penny Ross, and Jane Chandler, "Linked Data for Libraries: Benefits of a Conceptual Shift from Library-Specific Record Structures to RDF-based Data Models," 78th IFLA General Conference and Assembly (2012), http://www.ifla.org/past-wlic/2012/92-alemu-en.pdf.

	Deutsch Number	Instrument/ Voice	Instrument/ Voice	Instrument/ Voice
Auf den Sieg der Deutschen	81	voice	two violins	cello
Brüder, schrecklich brennt die Thräne	535	soprano	small orchestra	
Seligkeit	433	voice	piano	

Fig. 3. Deutsch number and instrumentation of three Schubert songs

model appears in various guises in various disciplines, for example: Node-Arc-Node (mathematics/graph theory), Subject-Predicate-Object (linguistics), Object-Attribute-Value (programming), Entity-Relationship-Value (software engineering), Record-Field-Data (relational database), Resource-Property-Value (information science). The RDF model can also be implemented in various forms. A good way to understand RDF is to recognize RDF in familiar places.

RDF READING OF A SPREADSHEET

Because each serves a distinct function, rows and columns of a spreadsheet are not interchangeable. For example, in a spreadsheet for instrumentation of musical pieces such as figure 3,[21] each row is a record about a piece of music, and each column represents the Deutsch number and an instrument used in the piece. The header of each row holds the title of the piece, and the header of each column designates what the information is about in the cells below. In an RDF reading of this table, title is the resource, Deutsch number and instrument/voice are the properties, and each cell contains the value. In other words, to construe a spreadsheet as RDF triples, the row header is the resource, the column header is the property, and the row-column intersection is the value, or: Row-Column-Cell, as shown in figure 4. In figure 5, I rewrote the spreadsheet as a set of RDF triples.

In essence, the structure of this particular spreadsheet can be configured as shown in figures 6 and 7. Note that the rows and columns with repeated headers need to appear only once in RDF, because RDF imposes no limits on the number of properties, including repeated ones, that a single resource can have. Figure 8 shows these relationships graphically.

21. For the purpose of illustrating contrasting data, I chose two Schubert songs in addition to *Seligkeit* for this and subsequent examples.

	Has Property: Deutsch Number	Has Property: Instrument/Voice	Has Property: Instrument/Voice	Has Property: Instrument/Voice
Resource: *Auf den Sieg der Deutschen*	Value: 81	Value: voice	Value: two violins	Value: cello
Resource: *Brüder, schrecklich brennt die Thräne*	Value: 535	Value: soprano	Value: small orchestra	Value: <empty>
Resource: *Seligkeit*	Value: 433	Value: voice	Value: piano	Value: <empty>

Fig. 4. RDF reading of a spreadsheet

Auf den Sieg der Deutschen	→ *has Deutsch Number*	→ 81
Auf den Sieg der Deutschen	→ *has instrument/voice*	→ voice
Auf den Sieg der Deutschen	→ *has instrument/voice*	→ two violins
Auf den Sieg der Deutschen	→ *has instrument/voice*	→ cello
Brüder, schrecklich brennt die Thräne	→ *has Deutsch Number*	→ 535
Brüder, schrecklich brennt die Thräne	→ *has instrument/voice*	→ soprano
Brüder, schrecklich brennt die Thräne	→ *has instrument/voice*	→ small orchestra
Seligkeit	→ *has Deutsch Number*	→ 433
Seligkeit	→ *has instrument/voice*	→ voice
Seligkeit	→ *has instrument/voice*	→ piano

Fig. 5. RDF triples of the spreadsheet in figure 3

While this spreadsheet can be construed as a set of RDF triples, the structure of this spreadsheet creates several constraints that limit the machine's ability to understand the data fully. The spreadsheet limits the number of entries for instrument/voice to three.[22] The three-column design compels data to be modified in certain situations. For *Auf den Sieg der Deutschen*, we enter "two violins" because entering "voice," "violin," "violin," "cello" requires four columns, so, to fit the data into three columns, the two appearances of "violin" are combined into a single entry "two violins." Allowing the use of the word "two," the meaning of the column is no longer unambiguous, because "two" is a number, not an instrument/voice. Moreover, the data is no longer atomic: because "two" and "violin" are two distinct pieces of data. For *Brüder, schrecklich brennt die Thräne*,

22. It might be easy to add another column in a spreadsheet application, but if this were a table as a part of a larger relational database, adding columns could be laborious. Altering the design of a relational database, such as adding a column, usually requires creating a development copy of the database and testing all existing functionalities against it.

	Deutsch Number	Instrument/Voice
Piece	number	name

Fig. 6. Data structure represented in the spreadsheet in figure 3

Piece $\rightarrow$ *has Deutsch number* $\rightarrow$ Number
Piece $\rightarrow$ *has instrument* $\rightarrow$ Name

Fig. 7. RDF reading of the spreadsheet in figure 6

Fig. 8. Graphic relationship of RDF elements in figure 7

there are not enough columns to list all the instruments of the orchestra, so we enter "small orchestra." In this case, the meaning of the column is, again, no longer unambiguous, because "orchestra" is an ensemble, not an instrument/voice, and "small" is a qualifier of the orchestra, not itself an instrument/voice.

Could we not change the column, then, to "instrument/voice or ensemble and the number thereof" so that we could capture as much information as possible in the limited space? While this appeals to human sensibility, machines would either be confused, or led to make inferences that are incorrect. On the other hand, changing the way we understand what instrumentation is about can lead to us structuring the data in a way that machines can understand.

CREATING MACHINE-PARSABLE DATA

Instrumentation, or medium of performance, is a complex concept. Decomposing the data in play reveals four components: part, instrument/ voice, player, and ensemble.[23] Illustrating them as RDF properties, these

23. Part, instrument/voice, player, and ensemble refer to the abstract concept, rather than the physical printed part, the physical instrument, the actual person, or a specific ensemble.

four components are interrelated as shown in the schematic in figure 9[24]: a piece of music consists of parts; each part calls for instruments/voices; each part also calls for certain types of players; each player is responsible for one or more parts; and various parts may be grouped into an ensemble.

This model resolves the atomicity and ambiguity problems we encountered earlier. If the score calls for two violins, as in *Auf den Sieg der Deutschen*, there will simply be two individual links to a violin part. If the score calls for a small orchestra, as in *Brüder, schrecklich brennt die Thräne*, there will be nine individual links to the nine orchestral parts, and then each of the nine parts will link out to a single orchestra.

This model can further resolve problematic situations toward describing medium of performance in current music cataloging practice. For example, instrumental doubling and generic instruments such as "percussion" can be expressed like this: a part is linked to multiple instruments; those instruments are all linked to one player; that player is linked back to the part. This level of specificity is possible because part, instrument/voice, and player are independent properties. Doing so also eliminates the need to enter the number of parts, the number of players, or the number of ensembles, because each of these numbers can be obtained by counting links, a task that machines can accomplish.

Other details of medium of performance can also be captured with more refined properties. For example, the "alternative medium of performance" concept (which is defined with subtle differences in MARC field 382 subfield p, and in UNIMARC field 146 indicator 2 and subfields b to f position 8) can be expressed using properties that signify alternative, and used only for the component in question. This leads to a more precise understanding of what is an alternative to what, in a number of distinct scenarios, including "same piece of music but consisting of different parts," and "same part but calling for different instrument/voice." Expressing alternatives this way not only covers situations where the alternative is explicit, such as a "sonata for clarinet or viola and piano," where the viola part is the alternative to the clarinet part while the piano part is unchanged. It also allows us to see other cataloging concepts in new light; for example, it is possible to express a piano/vocal version of an opera as the piano part being the alternative to all the orchestral

24. Earlier versions of this diagram with its technical underpinnings were presented at the 15 October 2016 meeting of the New York State-Ontario Chapter of the Music Library Association in Toronto, Canada; and on 8 July 2016 at the annual congress of the International Association of Music Libraries, Archives and Documentation Centres in Rome, Italy. I would like to thank my international colleagues for their valuable input.

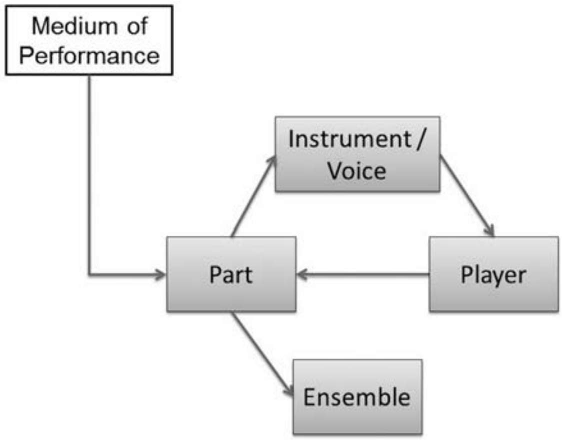


Fig. 9. Interrelation of four RDF components

parts together, while the voice parts remain unchanged. But we might want to make the distinction between these two types of alternativeness. With linked open data, we are free to refine the "alternative" property to a "derivative of" property. Or, refining to show various degrees and styles of derivativeness, such as "part adapted for" (another instrument), "orchestration of," "reduced orchestra version of," "piano reduction of," "re-orchestration of," "adapted for" (a different instrumentation), or even "reconstruction of," "re-creation of," "inspired by," "restyling of." While for a human user using "alternative" will suffice for all these scenarios, more precise properties allow machines to acquire more nuanced understanding, especially for complex concepts and the many degrees of equivalence and similarity.

For machines, it is perfectly acceptable to encounter relations that are not equivalent. Depending on the sophistication of the program, machines can do the job of analyzing the nature of the similarity, evaluating the degrees of similarity, and calculating the likelihood of usefulness when responding to a query, or the machine's version of taking a guess. So, the problem is not that medium of performance concepts are not equivalent between MARC field 382 and UNIMARC field 146, but is the lack of equivalence and similarity relationships defined to bridge the two. The same problem extends outside library data: no equivalence or similarity relationships exist for connecting library medium-of-performance data with other nonlibrary data sets, such as Discogs. To build a global web of linked open data, providing the means to connect them is key.

LIBRARY DATA AS LINKED DATA

The library community is fortunate to have quality data, created by trained specialists, in a uniform, structured database design. The downside is that as cataloging has evolved over time, idiosyncrasies have crept into our practice. Without knowing or realizing the full implication of linked open data technologies that would later emerge, we have inadvertently developed cataloging rules to accommodate data structures rather than atomic data and unambiguous properties, and we have modified data structures to accommodate conventional human usage and readability. These developments hinder machine-parsability, making it difficult for library data to be processed easily on the open Internet.

In recent years, however, the cataloging community has incrementally positioned itself to enable linked data implementations. Theoretical work and case studies have been done with the content standard RDA, the underlying conceptual model FRBR, and the future MARC format replacement, BIBFRAME.[25] For music, there are several active linked-data initiatives under way.[26] Nevertheless, it is extremely important to recognize that our cataloging practice has been focused on enabling human tasks. We operate on a set of looming assumptions that: (1) there is a thing (physical or electronic); (2) people are intentionally looking for it (or stumble upon it while looking for something else); (3) people want to get it into their possession (physically or electronically); and, (4) once in possession, they want to "use" it (to read, to play, to deploy, to somehow consume its content). By contrast, in the web of data, "people" make up a shrinking subset of the users, while machine processing is promoted. Programs and algorithms crawl the web of data to build knowledge of their own and to answer human queries. The questions for us today are how to supply data to this web of data,[27] and how to harness machines' analytical power for library users.[28]

---

25. For an explanation on modeling RDA in RDF, see Kimmy Szeto, "Positioning Library Data for the Semantic Web: Recent Developments in Resource Description," *Journal of Web Librarianship* 7, no. 3 (2013): 305–21; an analysis of FRBR and its applicability to the linked data environment, see Karen Coyle, "Bibliographic Description and the Semantic Web," in *FRBR Before and After: A Look at Our Bibliographic Models* (Chicago: ALA Editions, 2016), 137–56; a technical paper on modeling FRBR, RDA, and BIBFRAME and the tension between closed and open data can be found in Thomas Baker, Karen Coyle, and Sean Petiya, "Multi-Entity Models of Resource Description in the Semantic Web: A Comparison of FRBR, RDA, and BIBFRAME," *Library Hi Tech* 32, no. 4 (2014): 562–82.

26. For example, Linked Data for Production, "Performed Music Ontology," https://wiki.duraspace.org/display/LD4P/Performed+Music+Ontology; DOing REusable MUSical Data, http://www.doremus.org; the Europeana Data Model Documentation, http://pro.europeana.eu/page/edm-documentation; and the Music Notation Community Group of the World Wide Web Consortium, https://www.w3.org/community/music-notation.

27. A comprehensive overview of the linked data vision can be found in Tim Berners-Lee and Mark Fischetti, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor* (San Francisco: HarperSanFrancisco, 1999); and Tom Heath and Christian Bizer, *Linked Data: Evolving the Web into a Global Data Space* (San Rafael, CA: Morgan & Claypool, 2011).

28. Philip Schreur discusses how this paradigm shift affects library technical services in his article "The Academy Unbound: Linked Data as Revolution," *Library Resources & Technical Services* 56, no. 4 (2012): 227–37.

THE LINKED DATA PROMISE

As to our original search for the Schubert song, I can safely say linked open data could enable machines to overcome the uncertainties: Misspelled name? Google suggested the correct one. Wrong country? Geographic proximity would lead to singers from Switzerland assigned a higher likelihood. The song title not really ending with "*-lein*"? "*-keit*" would more likely be found as a partial match. Possibly a one-word song title? Short song titles would be given more weight. Possibly on the last track on the LP? Machines would understand tracks are often shuffled in reissues, thus giving this criterion less scrutiny. And, what about the high G? Software can now read and notate music with much improved accuracy, and G-sharp is in close proximity. The upbeat nature of the song? Proprietary online music streaming services have been developing algorithms to capture mood in music.

Linked open data invites us to reorient our approach to creating, managing, and curating data. In return, it lowers the barriers to accessing information and enables knowledge production on a massive scale. The technology is there, and we can, in fact, do better. But first, at least for the library community, we must do a better job working with machines so that machines can work better for us.

ABSTRACT

Linked open data promises global interconnectedness of a vast amount of data. Web technologies promise to lower the barriers to accessing information, and to enable knowledge production of massive scale. But can the web of data answer a music reference question? Starting with a seemingly impossible search for a Schubert song, this article describes how linked data technologies could overcome some limitations of catalog searching. Technical and conceptual challenges, however, are intertwined in the library community's effort to publish linked data. Through an analysis of contrasting data models, this article offers a linked data reading of medium of performance, and how the data can be tweaked to improve machine processing. This example leads to a discussion on general strategies toward an open, interoperable, evolvable, machine-actionable network that enables computers to become more effective tools for answering human questions.

Paper 4

Metadata standards in digital audio

in Khosrow-Pour, M. (ed.), *Encyclopedia of Information Science and Technology*. 4th ed. Hershey, Pennsylvania, USA: IGI Global, vol. 8, pp. 6447-6463.

DOI: 10.4018/978-1-5225-2255-3.ch560

a prior publication included in this dissertation,
redacted from the repository pending permission from the publisher

Encyclopedia of Information Science and Technology, Fourth Edition

Mehdi Khosrow-Pour, D.B.A.
Information Resources Management Association, USA

Volume VIII
Categories: MA - MU

IGI Global
DISSEMINATOR OF KNOWLEDGE
www.igi-global.com

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Khosrow-Pour, Mehdi, 1951- editor.
Title: Encyclopedia of information science and technology / Mehdi Khosrow-Pour, editor.
Description: Fourth edition. | Hershey, PA : Information Science Reference, [2018] | Includes bibliographical references and index.
Identifiers: LCCN 2017000834| ISBN 9781522522553 (set : hardcover) | ISBN 9781522522560 (ebook)
Subjects: LCSH: Information science--Encyclopedias. | Information technology--Encyclopedias.
Classification: LCC Z1006 .E566 2018 | DDC 020.3--dc23 LC record available at https://lccn.loc.gov/2017000834

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

For electronic access to this publication, please contact: eresources@igi-global.com.

Metadata Standards in Digital Audio

M

Kimmy Szeto
Baruch College, City University of New York, USA

[redacted]

DOI: 10.4018/978-1-5225-2255-3.ch560

[redacted]

6448

[redacted]

6449

[redacted]

6450

[redacted]

6451

[redacted]

6452

[redacted]

6453

[redacted]

6454

[redacted]

[redacted]

6456

[redacted]

[redacted]

6458

[redacted]

6459

[redacted]

6460

[redacted]

6461

[redacted]

6462

[redacted]

6463

Paper 5

The roles of academic libraries in shaping music publishing in the digital age

This article first appeared in
Library Trends, 67(2), Fall 2018, pp. 303-318

DOI: 10.1353/lib.2018.0038

a prior publication included in this dissertation per publishing agreement (Use 2)

The Roles of Academic Libraries in Shaping Music Publishing in the Digital Age

Kimmy Szeto

Abstract
Libraries are positioned at the nexus of creative production, music publishing, performance, and research. The academic library community has the potential to play an influential leadership role in shaping the music publishing life cycle, making scores more readily discoverable and accessible, and establishing itself as a force that empowers a wide range of creativity and scholarship. Yet the music publishing industry has been slow to capitalize on the digital market, and academic libraries have been slow to integrate electronic music scores into their collections. In this paper, I will discuss the historical, technical, and human factors that have contributed to this moment, and the critical next steps the academic library community can take in response to the booming digital music publishing market to make a lasting impact through setting technological standards and best practices, developing education in these technologies and related intellectual property issues, and becoming an active partner in digital music publishing and in innovative research and creative possibilities.

Introduction
Academic libraries have been slow to integrate electronic music scores into their collections even though electronic resources are considered integral to library services. The Association of College and Research Libraries considers electronic resources integral to information literacy, access to research, and collection policies in academic libraries (ACRL 2006a, 2006b). Collection development surveys conducted by the National Center for Educational Statistics indicate electronic books, database subscriptions, and electronic reference materials constitute roughly half the materials

LIBRARY TRENDS, Vol. 67, No. 2, 2018 ("The Role and Impact of Commercial and Noncommercial Publishers in Scholarly Publishing on Academic Libraries," edited by Lewis G. Liu), pp. 303–18.

budget in academic libraries in the United States (2012). While librarians continue to examine the impact of electronic books and databases (Lamagna, Hartman-Caverly, and Swenson Danowitz 2015; Walters 2013; Walters 2014; Durant and Horava 2015), literature on electronic musical scores has been more scarce, reflecting the format's state of integration into library collections. With electronic sheet music publishing on the rise (McGinley 2016b), libraries are positioned at the nexus of creative production, publishing, performance, and research. In this paper, I will discuss the factors that have contributed to this moment, and the critical next steps the academic library community can take to become an influential player, together with music publishers, in the electronic scores ecosystem.

Electronic Scores in Libraries

Library offerings of electronic scores are generally offered in the same way as electronic books or online databases even though the nature of the demands and uses for music scores differs significantly. Ana Dubnjakovic (2009) described the recent proliferation of digitized sheet music online and offered advice on evaluating the quality of the sources and effective searching. Lisa Hooper (2013) issued a call-to-action to initiate a "dialogue between music librarians, vendors, publishers, acquisition librarians, and other non-music librarian professionals" (575). Yet libraries have continued to be reactive to the evolving publishing landscape and complementary technologies. In his 2015 speech, when comparing electronic scores to digital text and the burgeoning field of digital humanities, Darwin Scott declared the state of electronic score "fractured, stuck in nascent and divergent stages of development." He described libraries as meeting the electronic score format in a relationship that is "murky and sometimes stormy," and the effort to integrate it into library operations as "bumpy," resulting in "collective frustration" and "passive surrender." Scott's sentiment was echoed in Hooper's presentation (2015) where she called for positive action and advocacy, with a focus on influencing publishers and vendors on pricing models, licensing models, user interface, and cataloging.

Since then, the music library community has responded with several technical responses. Acknowledging the prevalence and importance of self-publishing, Kent Underwood produced a landmark study of contemporary composers self-publishing works online (2016), and established an archival process for these composers' websites with curator Robin Preiss (Underwood and Preiss, n.d.). Reed David and Nurhak Tuncer (2016, 2017) found significant music publishing activities occur online in their survey of bibliographic cataloging practices of self-published items. They discovered that most bibliographic records of this type originated from academic libraries, which suggests that academic libraries represent the site of significant collecting activity. Meanwhile, Adams and Levy (2017) questioned the nature of publishing itself by way of focusing their study on the bibliographic cataloging of print-on-demand scores in physical format. They recognized that digital engraving and printing technologies have given rise to a print-on-demand industry, which blurs the distinct cataloging concepts of publication, distribution, and printing. Such blurring and merging of publishing processes will put academic libraries in a unique position to influence the entire life cycle. Behind the issues of cataloging is the peculiar collection practice where scores are acquired in the electronic format but then the library prints, binds, and circulates them as physical copies, as discussed in Peters (2017).

State of Electronic Scores

Attempting to define the term "electronic score" or "digital score" is a fraught enterprise. The perils stem from the wide applicability of the concept "music score," and the divergent, nonunified situation of its electronic format. In particular, the influence of electronic books, which are not analogous to electronic scores, may hinder emergent best practices, in that applying the considerations for one to the other will make it more difficult for libraries to integrate electronic scores as a collection category.

On the surface, the digital image makes a serviceable electronic counterpart to the physical page. However, such an image is a static, final form document, an image, which, from the perspective of a computer, is no different from any other graphic or digital photo, containing no actionable data other than pixels of color. Today, with digital text, one can reasonably expect to be able to perform some dynamic functions on the text, such as searching, highlighting, adaptive display, reading aloud, and in certain circumstances, light editing. These functions are made possible by the presence of the textual content embedded in a digital text underlay, where each letter and punctuation is as if "typed up" and linked to its corresponding location on the image. This machine-actionable layer can be created natively from the start when a document originates from a word processor, which subsequently generates the image. Or, the digital image can be created first, such as by scanning, and then the computer "reads" the image through optical character recognition (OCR) software, which generates the digital text underlay. This image-cum-text package is commonly distributed via the Portable Document Format (PDF), popularized by the Acrobat software, which serves primarily as the OCR software and screen reader.

Electronic scores have also been produced and distributed widely via PDF. However, the distribution of music notational content is complicated by a few factors. First, no dominant, open standard has yet to emerge to support the encoding of the variety of musical notation systems and fonts. As a result, the widely and freely available PDF infrastructure remains the de facto standard for the music score, but is capable of providing viewing only, without any dynamic functions. Second, improving the quality of optical music recognition (OMR), encoding music notation through reading digital images of printed music, continues to be a technological challenge (Rebelo et al. 2012). Third, the visual formatting of music notation for printing or screen display, called engraving, is so complex that computers are only beginning to be able to produce satisfactory layouts comparable to those created by hand.

Advances in computing power and machine learning are enabling great strides in OMR and digital engraving. The emergence of open standards such as MEI, MusicXML, and SMuFL have empowered digital scholarship and facilitated the exchange of music notation files across music writers. As a result, the creation and distribution of natively created digital scores as well as score reader software that offers dynamic functions has increased.

Music Publishing

Music Publishing

Music publishing distinguishes itself from general publishing as a subset that specializes in issuing products that consist primarily of musical notation (Krummel 2001). George Sturm (2000) characterized music publishing as "the *art* of <u>bringing</u> a *musical product* to <u>a</u> *public*" (628; emphasis in original, underlining added). This statement warrants unpacking. First, even though the artistic work is communicated through sound, the product exists, traditionally, in the form of a score, and producing a musical score is itself undeniably an art. Second, the intended audience of this product is not *the* general public, but rather, *a* certain public, a narrower demographic of users and collectors who tend to have specific needs. Finally, music publishers have various ways in which to *bring* out their products.

For libraries, it is important to understand certain idiosyncratic behaviors of music publishers. The traditional roles of music publishers are as follows: commissioning, financing the production, promoting the publications, and distributing the product. Publishers serve as the mediator between supply and demand: they seek out promising composers and songwriters and find or create a market for their works. In so doing, a publisher may purposely withhold a score at any of the following stages: production, publication, distribution, reproduction. Krummel (2001) put these behaviors into three categories: produced but not published, such as luxury editions privately commissioned by a collector; printed but not published, when the composer wants to directly negotiate royalties and control the performances; published but not printed or distributed, when manuscripts and handwritten copies sufficed or are preferred.

This last category presents a particular complication. Manuscript scores are preferred when the score is expected to undergo frequent changes, such as in an operatic score, and when the composer intends to depict ideas beyond the capabilities of standard notation and engraving. The tradition of copying scores (by hand) is so ingrained in music publishing that the manuscript score still occupies a good portion of publications, and the notational contents in these scores are the most technologically challenging to transfer to digital form.

Music Printing

The history of music publishing is interwoven with the history of technology, society, and commerce. Musical notation has been used for memory aid for millennia, and, in the past three hundred years or so, became a separate artifact of a musical "work" apart from the act of performing the music. Producing musical notation is difficult. Writing out the notation requires a deep knowledge of the musical content, as well as discipline and precision to the spacing and graphic details of the symbols. Printing techniques using woodblocks, stones (lithography), and metal plates (intaglio) require not only music literacy and penmanship but also craftsmanship to produce the reverse negative and ink the medium. Also, these mediums are not reusable. Movable, reusable type presents printers with a different dilemma: invest in a very large amount of type to account for every possible overlapping of musical elements and print with a single impression, or run the same sheet of paper through multiple impressions of separate musical elements but risk misalignment and waste of paper and ink. Eventually, photographic methods rendered the dilemma moot, and photoengraving on stone, copper, and zinc plates became the dominant method.

The liberation of printing from the limitation of printable symbols encouraged composers to experiment beyond conventional standard notation. However, music publishers producing scores by computer run up against the same constraints as they did prior to the development of the photoengraving method, as score writers are limited to symbols available in the font and the engraving capability of the software.

Music Publishing for Academic Purposes

Until the mid-nineteenth century, consumers of music scores were primarily performers. Accordingly, music used to be sold in music shops together with music supplies, rather than in book shops. The business changed when public concerts and music literacy became more widespread. This period also saw the rise of the academic study of music performance and musicology, which led to the demand and production of facsimile editions of manuscripts, sometimes with their corresponding standard notation counterpart, and reproductions of early editions of published music. Public concerts and sound recordings also increased interest in the general public for study scores, which are printed in a smaller size, often accompanied by commentaries, analytic essays, and corrections to the original score.

In addition, academic libraries also tend to collect the following types of scores: performance editions, which are laid out in a larger size with convenient page turns, often with commentaries and performance guidance; scholarly editions, which are edited to reflect the historical and bibliographical study of the provenance and transmission of works; and complete editions, which represent all the output of a particular composer. The impulse to anthologize is particularly strong in the United States due to a period of active reprinting of repertoire lost in Europe to the Second World War. However, because many of the original plates are lost, such republication efforts were often accomplished by photographic or photolithographic reproductions from earlier printed copies, with varying results.

Sheet Music Publishing

Generally speaking, the classical repertoire is continually being copied and re-engraved digitally, and new compositions are being created digitally. However, libraries continue to collect physical editions, and consumers, even when purchasing online, continue to find music publishers offering only static digital images of physical editions. Meanwhile, the sheet music publishing industry overall has been steadily declining, and has been slow to capitalize on the growing digital market.

Industry reports on the ten-year period beginning in 2006 in the United States show that sheet music publishing overall (print and digital) experienced a 40% decline (McGinley 2016b), while book publishing industry revenue decreased 17% in print (Rivera 2017), but was accompanied by an one hundred-fold increase in the digital market (McGinley 2016a). The analyst who reported on the sheet music industry attributed the current decline to dwindling demand due to declining music education, literacy, and piracy, but maintained a positive five-year outlook comparable to electronic book publishing due to growth in the digital market.[1] At present, for the sheet music industry, digital publishing is akin to selling the same products, while eliminating the costs for printing and distribution, as well as for the metal plating and lithographic engraving process.

It is true that music publishers have begun selling scores electronically. More precisely, music publishers have adopted the business model to sell licenses to digital copies of static electronic scores. The licensing model affords music publishers the ability to sell their products in smaller units and maintain more control through placing restrictions on usage, such as the time period in which the score is accessible and the number of times the score can be printed or circulated. These restrictions are accomplished through allowing access via subscriptions and via proprietary software. The end result for the consumer is essentially the same as purchasing print, that is, to obtain a digital substitute for the print, without most of capabilities of the digital medium.[2]

Music publishers have just begun to take advantage of the potential dynamic functions of the digital medium. The reasons are not solely due to lowering cost and increasing revenue. The shadow of electronic books, technological constraints, and user behaviors all contribute to the delayed electronic boom.

Technology of Electronic Scores

Encoding of Music Notation

Encoding text is relatively simple because text is a one-dimensional sequence of letters. Music notation, on the other hand, involves capturing multiple streams of symbols that vary in length and size and interact with symbols in other streams in different ways depending on context. Encoding music notation into structured machine-readable and machine-actionable data is even more complex, since the notation's context dependency cannot be translated to simple rules.

This complexity is especially true for engraving, where, over three decades of development, software is still unable to achieve the level of visual clarity expected from hand-engraved scores. The challenge with the encoding is, therefore, to balance the amount of notational data that needs to be encoded, the complexity of the rendering or analytical engine, the computing power required, and the amount of human intervention that is expected. This challenge has contributed to the high cost of music notation software relative to text, and the long period of development before the technology becomes widely affordable.

Encoding Standards

There have been many players in the development of structural representation of music notation. Early development for standard western notation has been dominated by two commercial score writers: Finale (first released in 1988) and Sibelius (first released in 1996). Both software programs focused on the graphical production environment: replacing the manual music notation input and engraving, and generating printed scores and static PDF. The proprietary nature of these two software programs kept the cost high and prevented development of compatible score readers that offered dynamic functions. Meanwhile, open-source developments progressed at a slower pace. The open-source software LilyPond was first released in 1998, but this text-based, engraving-focused engine has not enjoyed widespread use. The Humdrum format has long been used for music analysis since the 1980s, but its software tools were purposely not developed to include engraving or printing (Huron 2001).

In recent years, open standards, especially MEI and MusicXML, have matured enough to compete with the two proprietary standards. The Music Encoding Initiative (MEI) began as a joint project of the University of Virginia Library and Der Akademie der Wissenschaften und Literatur in Mainz, and has now turned into a community-driven collaboration. MEI is geared toward scholarly publishing of historical repertoire, including encoding of musical notation from the Medieval and Renaissance periods, as well as music-analytical, historical, and bibliographic information (Hankinson, Roland, and Fujinaga 2011).

MEI compares itself to MusicXML as the format that focuses on capturing intellectual content in an existing physical musical document, while MusicXML is designed to mediate between commercial score writers (MEI 2018). First released in 2004, MusicXML, originating from a software company, was developed to facilitate exchange of files between Finale and Sibelius (initially via plugins, although now conversion is supported natively in both software). In 2015, development was transferred to the W3C Community Group, an open web platform currently dominated by software companies and music publishers (see the list of persons and institutions who have committed to the Final Specification Agreement in Music Notation Community Group 2017). In the same year, the open-source score writer MuseScore, which supports MusicXML, released a major version update that offers functionalities comparable to those offered by the other commercial software. Currently, work is in development to better integrate MusicXML to the open music notation font platform SMuFL (Good 2017).

Encoding tablature and chord notations is relatively simpler, and the situation is less fraught. The widely used commercial proprietary format and editing software Guitar Pro is well documented (Vromman, n.d.) and has been adopted by other software programs. Free, open-source alternatives such as PowerTab and ChordPro provide similar functionalities. These three formats are found being implemented on websites, such as Ultimate Guitar and Chordie, to perform dynamic display and transpositions, which offer a glimpse of dynamic functions possible in electronic scores.

The convergence of open encoding standards and software has spurred the production and use of electronic scores. The earlier focus on developing the production environment is now shifting to the user side, especially since tablet devices have become more affordable. With open standards, electronic score readers have proliferated and are gradually adding dynamic functions to music scores. These score viewers are annotation tools for PDF files geared toward music scores, and are capable of manipulating encoded music scores, such as on-the-fly re-engraving in response to screen size and user resizing, and interactive functions such as selective display, selective playback, transposition, search and highlight, and user annotation (Winget 2008). In addition, score readers now commonly have the ability to communicate with other devices for exchanging files, page turning with external pedals, and synchronized controls, as well as interfacing with online storage (Szeto 2018). However, this technology stack is not necessarily standard in library services or in a music information literacy curriculum.

Users of Electronic Scores

Use of Music Scores

Music notation serves a dual ontology: to depict musical ideas, which, in turn instruct the performance of the sounds. From the user perspective, there is a large difference between reading a book and reading a music score, which often necessitates more extensive scrutiny of the notation and examination of multiple publications of the same work.

In realizing pitch, duration, and sometimes words, the process is meant to be carried out continuously in time, and, in works with more than one part, to be synchronized across multiple performers. Depending on circumstance, users may seek the score, which presents all the parts in a single view, or the part, which presents only a single performer's view. In vocal music, a user might be interested in seeing the score written in a transposition, that is, essentially the same piece of music in a different key or at a different overall pitch level. In print, scores, parts, and transpositions are all published as separate items. Library cataloging practices traditionally address these issues at the FRBR expression level. Rules for distinguishing scores from parts have long been in place, but musical key has not been a consistent element of description until recently, an omission from the catalog that has historically presented difficulties for library users.

Users also seek multiple editions for their physical layout and accompanying contents. Music scores, especially performance editions, are meant to be read by performers some distance away from the score so that individual preference for the font and size of notation, spacing, and pagination factors strongly. Users also seek to compare editorial annotations—instructions, guidance, suggestions, and translations added to the musical notation—which are useful for realizing the score as well as for historical performance study (R. Scott 2013).

Lastly, the reading of a music score requires extensive personal interaction, which is a contributing factor to the "music score's slow entrée into the digital realm," as Hooper (2015) describes:

> We analyze it, we write in our own fingerings, we change bowings, we add other visual cues, and all of this we write directly into the score. We do this because looking at a musical score is rarely a one-off occurrence. We read and reread a score from start to finish a thousand times, dissect it into the tiniest fragments until the music is ingrained in our mind, in our fingers, and, I am sure some would say, in our souls. In short, a student musician's typical interaction with a score is far more active than a student's typical interaction with a book. (571)

Characteristics of Electronic Scores

In fact, the particular nature of use of music scores actually lends itself very well to the digital form. Dynamic electronic scores can provide the following:

- Portability:
  - can be accessed anywhere without having to obtain and carry multiple editions
  - can be displayed on any device
- Legibility:
  - can adapt to different screen sizes
  - can adjust the display size (font size)
  - can accommodate legible and musically sensible annotation and highlights
- Manipulability of musical content:
  - can be displayed or played back in transpositions
  - can be played back with custom instrumentation
  - can be displayed or played back selectively
  - can be played back at custom speeds
  - can be searched and evaluated quantitatively

Music publishers are beginning to see the demands for these dynamic functions and are now beginning to offer services that include display, playback, and transposition options in mobile applications.

Electronic Scores on Cloud Platforms

Cloud applications have made possible a single destination for creation, display, and interaction of scores. In this computing model, contents and functionalities are accessed online and are selectively served to users based on permissions, thus eliminating the need for separate, locally installed score writer and score reader applications. When a user creates a score using a cloud application, the score is automatically stored in the cloud. The creator can then grant permissions to other users so that they can access certain dynamic functions offered by the application.

Music publishers and software developers alike are now migrating to the cloud platform. Noteflight, initially developed by textbook publisher W.W. Norton as an online interactive music theory workbook, has evolved into a general music notation tool with emphasis on the education market. Sibelius followed with the service Sibelius Cloud Publishing, which is aimed toward composers and arrangers for creating, publishing, and marketing their works. Meanwhile, the music publisher and distributor Hal Leonard acquired Noteflight (in 2014) and the online retailer Sheet Music Plus (in 2017) to create a cloud platform for selling and licensing existing works.

All these transformations might be transparent to the average user, who will see their production environment and retail experience largely unchanged. However, this technological leap will revolutionize the music publishing industry because the platform liberates electronic scores from PDF to exploit the dynamic functions, and at the same time erases the boundary between production, publishing, and consumption.

Academic Libraries in Electronic Score Publishing

Academic Support

Academic libraries' involvement in this electronic score revolution has been limited largely to MEI, which has demonstrated libraries' unique position to breathe new life to old music, especially the medieval and Renaissance corpora, which are getting much attention since computational and empirical music research becomes possible. Structured encoding such as MEI, MusicXML, and Humdrum, together with programming tools like Python and music21, enable searching musical elements and notational symbols within musical works. Much like full-text searching in digital humanities applications, the ability not only gives library users the ability to search beyond catalog descriptions, it can also enable novel analytical methodologies that generate new musicological insights and perspectives (see Fujinaga, Hankinson, and Cumming 2014). A cursory search on how computers have played a role in music scholarship quickly returned a wealth of recent research that has uncovered new voice-leading principles through computer analysis of fifteenth- and sixteenth-century music theory treatises (Morgan 2016); furthered the application of neural networks to analyzing musical structures (De Valk 2015) and orthography, which improves OMR (Cherla 2017); spurred the development of music informatics (Steyn 2013) and big music data infrastructure (Fournier-S'niehotta, Rigaux, and Travers 2016; Abdallah et al. 2017); facilitated new digital methods of cultural studies (Serra 2017); and advanced human-computer interactive performances (Delgado, Fajardo, and Molina-Solana 2011; Kirke and Miranda 2013).

Beyond collecting and guiding users to online collections of electronic scores, libraries can play a systematic, strategically active role in the development of the electronic score infrastructure and integrating related services. Many academic libraries already support electronic text by providing scanners, content management, and software tools. Supporting electronic scores is very similar, perhaps with the addition of circulating tablets and page-turning peripherals. With vertically integrated services in place, the library is essentially also supplying the necessary technology for electronic score publishing, as well as a means for bibliographers to develop and maintain collections of electronic scores.

Libraries can even go one step further, perhaps consortially, to establish digital library platforms for music research (for example, as demonstrated in Arora 2011), and build the corpus by encoding their own score holdings, especially the vast repertoire of music in the public domain. In fact, libraries can take the lead role in educational policy—should this technology stack become a standard library service—to institute the technological component in the information literacy curriculum for music students, just as online research and word processor skills are in a general curriculum.

Another leading role libraries can play involves the proper management of intellectual property when offering electronic score publishing to library users. Copyright issues have historically been a complex issue in music publishing (Meek 1953) and encompass various types of rights (moral, literary, performance, mechanical, and grand). The applicability and nonapplicability of the fair use doctrine and the first sale doctrine add complexity to licensing and ownership. Libraries, possibly bringing together expertise from an institution's legal department, can tackle these issues on several fronts. They can educate library users on what can lawfully and unlawfully be done with an electronic score. They can educate composers and arrangers on the circumstances under which copyrighted music can be reused and incorporated in their own works, as well as how to protect their own copyright as they publish their works. These efforts will involve participating in developing new standard licenses; interacting with collective rights management, such as with performance rights organizations; and, more generally, educating library users to think more critically about the economic, social, cultural, and legal ramifications of their intellectual rights, since now libraries are providing support as a publishing platform (Brown and Waelde 2018).

Technical Support

Libraries, especially academic libraries, interface between a population of heavy music users, creators of sheet music, public domain sheet music contributors, and music publishers. This is a unique position from which to lead the publishing industry and to standardize and simplify electronic scores. The result, ultimately, will shape the entire ecosystem so that electronic scores flow more smoothly and become more useful.

Historically, libraries have already proved to be influential in setting technical standards for information systems. As the commercial publishing market begins to converge on the open standard MusicXML, libraries really ought to reach out and weigh in. A music encoding standard that works well together with emerging bibliographic standards such as Bibliographic Framework Initiative and Performed Music Ontology will be the key to integrating electronic scores workflow through the entire life cycle of music scores.

Standardizing the production chain goes beyond standardizing encoding—it also includes engraving specifications, embedded metadata, and embedded rights management, all of which will facilitate music distribution for creators. On the user side, libraries can then use these technical production standards as a basis for discovery system requirements. Compared to the current requirements, which are based on descriptive metadata (Newcomer et al. 2013; Music Discovery Requirements Update Task Force 2017), the inclusion of semantic, rights, and other types of embedded metadata can make searching and automated processes much more powerful.

Creative Support

On the creative side, academic libraries can become centers or major supporters of research in composition and performance. In real-time notation, also variably called live notation, dynamic notation, live coding, live scoring, virtual scoring, and reactive notation, the "score" of this performance art is continually created and transformed as a response to the very performance itself as it takes place (Clay and Freeman 2010). These "scores" can even be networked so that each ensemble musician is served an individually tailored part on their mobile device in real time (Eldridge, Hughes, and Kiefer 2016; Onttonen 2017). Another type of performance art, "augmented musical scores," connects data streams—video, audio and motion sensors, the performers' sounds and biomechanical movements—turning live transactions of all kinds into a multimedia, human-computer performance (Tanaka 2000; Hope 2017). Other creative areas that can take advantage of electronic scores involve audience-assisted composition (Freeman 2008) and computer-assisted composition, where music notation intersects with computer programming in a composition environment such as OpenMusic (Agon 1998; Agon, Assayag, and Bresson 2018).

Conclusion

Academic libraries thus have the potential to become the force that empowers a wide range of creativity and scholarship: music performance, musicology, typography, cultural studies, performance studies, and human-computer interaction. In the publishing marketplace, academic libraries can play a leading role in shaping the industry life cycle so that new products are more readily discoverable and accessible, which will ultimately benefit music publishers as well as creators and consumers. Academic libraries are poised to take the lead and make a lasting impact on setting technological standards and best practices, developing education in these technologies and related intellectual property issues, and becoming an active partner in innovative creative possibilities.

Notes

1. Print book industry revenue declined from $32.2 billion in 2006 to $26.67 billion in 2016, with a five-year outlook of 2.2% annualized growth (Rivera 2017); electronic book industry revenue grew from under $35 million in 2006 to $3.8 billion in 2016, with a five-year outlook of 1.3% annualized growth (McGinley 2016a); sheet music industry revenue declined

from $457.7 million in 2006 to $268.5 million in 2016, with a five-year outlook of 1.1% annualized growth (McGinley 2016b). These figures reflect only the sales of publications, excluding revenue from various forms of licensing and royalties (see calculation by analyst Will Page in Ingham 2016).

2. From the FAQs on three major online sheet music retailers—musicnotes.com, onlinesheetmusic.com and sheetmusicplus.com—all allow one transposition and one print per purchase and unlimited viewing. The first two sites use a proprietary viewer, while the third places printing restrictions within the PDF file.

References

Abdallah, Samer, Emmanouil Benetos, Nicolas Gold, Steven Hargreaves, Tillman Weyde, and Daniel Wolff. 2017. "The Digital Music Lab: A Big Data Infrastructure for Digital Musicology." *Journal on Computing and Cultural Heritage* 10 (1): 1–21. https://doi.org/10.1145/2983918.

ACRL (Association of College and Research Libraries). 2006a. "ACRL/SAA Joint Statement on Access to Research Materials in Archives and Special Collections Libraries." American Library Association. September 6, 2006. http://www.ala.org/acrl/standards/jointstatement.

ACRL (Association of College and Research Libraries). 2006b. "Guidelines for University Library Services to Undergraduate Students." American Library Association. September 1, 2006. http://www.ala.org/acrl/standards/ulsundergraduate.

Adams, Anne, and Morris Levy. 2017. "Cataloging Scores in an Age of Print on Demand." Paper presented at the Annual Meeting of the Music OCLC User Group, Orlando, Florida, February 21–22, 2017. http://musicoclcusers.org/wp-content/uploads/AdamsLevy.pdf.

Agon, Carlos. 1998. "OpenMusic: Un langage visuel pour la composition musicale assistée par ordinateur." PhD diss., Université Pierre et Marie Curie.

Agon, Carlos, Gérard Assayag, and Jean Bresson. 2018. "OpenMusic." Institute for Research and Coordination in Acoustics/Music. Last modified March 13, 2018. http://repmus.ircam.fr/openmusic/home.

Arora, Nitin. 2011. "MXMLiszt: A Preliminary MusicXML Digital Library Platform Built on Available Open-Source Technologies." *OCLC Systems & Services* 27 (4): 298–316.

Brown, Abbe E. L., and Charlotte Waelde, eds. 2018. *Research Handbook on Intellectual Property and Creative Industries.* Northampton, MA: Edward Elgar.

Cherla, Srikanth. 2017. "Neural Probabilistic Models for Melody Prediction, Sequence Labelling and Classification." PhD diss., City University of London.

Clay, Arthur, and Jason Freeman. 2010. "Virtual Scores and Real-Time Playing." *Contemporary Music Review* 29 (1): 1. https://doi.org/10.1080/07494467.2010.509587.

David, Reed, and Nurhak Tuncer. 2016. "The Cataloging of Self-Published Items." Paper presented at the Annual Meeting of the Music OCLC User Group, Cincinnati, Ohio, March 1–2, 2016. http://musicoclcusers.org/wp-content/uploads/2016MOUG_TuncerDavid.pptx.

———. 2017. "The Cataloging of Self-Published Scores and Recordings." Paper presented at the Annual Meeting of the Music OCLC User Group, Orlando, Florida, February 21–22, 2017. https://goo.gl/FM8CxP.

De Valk, Reinier. 2015. "Structuring Lute Tablature and MIDI Data: Machine Learning Models for Voice Separation in Symbolic Music Representations." PhD diss., City University of London.

Delgado, Migual, Waldo Fajardo, and Miguel Molina-Solana. 2011. "A State of the Art on Computational Music Performance." *Expert Systems with Applications* 38 (1): 155–60. https://doi.org/10.1016/j.eswa.2010.06.033.

Dubnjakovic, Ana. 2009. "Navigating Digital Sheet Music on the Web: Challenges and Opportunities." *Music Reference Services Quarterly* 12 (3): 3–15. https://doi.org/10.1080/10588160902894972.

Durant, David M., and Tony Horava. 2015. "The Future of Reading and Academic Libraries." *portal: Libraries and the Academy* 15 (1): 5–27. https://muse.jhu.edu/article/566420.

Eldridge, Alice, Ed Hughes, and Chris Kiefer. 2016. "Designing Dynamic Networked Scores to Enhance the Experience of Ensemble Music Making." In *Proceedings of Second International Conference on Technologies for Music Notation and Representation*, edited by Richard Hoadley, Chris Nash, and Dominique Fober, 193–99. East Anglia, UK: Anglia Ruskin University Press.

Fournier-S'niehotta, Raphaël, Philippe Rigaux, and Nicolas Travers. 2016. "Is There a Data Model in Music Notation?" In *Proceedings of Second International Conference on Technologies for Music Notation and Representation*, edited by Richard Hoadley, Chris Nash, and Dominique Fober, 1–6. East Anglia, UK: Anglia Ruskin University Press.

Freeman, Jason. 2008. "Extreme Sight-Reading, Mediated Expression, and Audience Participation: Real-Time Music Notation in Live Performance." *Computer Music Journal* 32 (3): 25–41. https://doi.org/10.1162/comj.2008.32.3.25.

Fujinaga, Ichiro, Andrew Hankinson, and Julie Cumming. 2014. "Introduction to SIMSSA (Single Interface for Music Score Searching and Analysis)." In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, 1–3. New York: ACM. https://doi.org/10.1145/2660168.2660184.

Good, Michael, ed. 2017. *MusicXML Version 3.1: Final Community Group Report.* W3C Music Notation Community Group. December 7, 2017. https://www.w3.org/2017/12/musicxml31/.

Hankinson, Andrew, Perry Roland, and Ichiro Fujinaga. 2011. "The Music Encoding Initiative as a Document-Encoding Framework." In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, edited by Anssi Klapuri and Colby Leider, 293–98. Miami, FL: University of Miami Press. http://ismir2011.ismir.net/papers/OS3–1.pdf.

Hooper, Lisa. 2013. "Contemplating E-Scores: Open Ruminations on the E-Score, the Patron, the Library, and the Publisher." In *Proceedings of the Charleston Library Conference,* edited by Beth R. Bernhardt, Leah H. Hinds, and Katina P. Strauch, 571–75. West Lafayette, IN: Purdue University Press. https://doi.org/10.5703/1288284315326.

———. 2015. "The Direction Things are Heading, the Place We Don't Want to End Up." Paper presented at the Annual Meeting of the Music Library Association, Denver, Colorado, February 25-March 1, 2015.

Hope, Cat. 2017. "Electronic Scores for Music: The Possibilities of Animated Notation." *Computer Music Journal* 41 (3): 21–35. https://doi.org/10.1162/COMJ_a_00427.

Huron, David. 2001. "Humdrum: Frequently Asked Questions." Last modified July 9, 2001. https://csml.som.ohio-state.edu/Humdrum/FAQ.html.

Ingham, Tim. 2016. "The Global Music Copyright Business Is Worth More than You Think, and Grew by Nearly $1bn Last Year." *Music Business Worldwide*, December 13, 2016. https://www.musicbusinessworldwide.com/the-global-music-copyright-business-is-worth-more-grew-nearly-1bn-last-year/.

Kirke, Alexis, and Eduardo R. Miranda, eds. 2013. *Guide to Computing for Expressive Music Performance.* London: Springer. https://doi.org/10.1007/978-1-4471-4123-5.

Krummel, Donald W. 2001. "Publishing," section 2 in "Printing and Publishing of Music." *Grove Music Online.* http://http://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040101.

Lamagna, Michael, Sarah Hartman-Caverly, and Erica Swenson Danowitz. 2015. "Integrating E-Books into Academic Libraries: A Literature Review." *Internet Reference Services Quarterly* 20 (1–2): 19–32.

McGinley, Devin. 2016a. *E-Book Publishing in the US.* IBISWorld Industry Report OD4579. December 2016. http://clients1.ibisworld.com/reports/us/industry/default.aspx?entid=4579.

———. 2016b. *Sheet Music Publishing in the US.* IBISWorld Industry Report OD4845. November 2016. http://clients1.ibisworld.com/reports/us/industry/default.aspx?entid=4845.

Meek, Marcellus R. 1953. "International Copyright and Musical Compositions." *DePaul Law Review* 3 (1): 52–81.

MEI (Music Encoding Initiative). 2018. "How Is MEI Different from MusicXML?" MEI (website). http://music-encoding.org/about/.

Morgan, Alexander. 2016. "Renaissance Interval-Succession Theory: Treatises and Analysis." PhD diss., McGill University.

Music Discovery Requirements Update Task Force, Music Library Association. 2017. *Music Discovery Requirements II.* October 2017. http://www.musiclibraryassoc.org/mpage/mdr_es.

Music Notation Community Group. 2017. "Final Specification Licensing Commitments for MusicXML Version 3.1." Last modified December 19, 2017. https://www.w3.org/community/music-notation/spec/137/commitments.

National Center for Educational Statistics. 2012. "Supplemental Academic Libraries Survey (ALS) 2012 Tables to *NCES 2014-038.*" https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014038.

Newcomer, Nara L., Rebecca Belford, Deb Kulczak, and Kimmy Szeto, with Jennifer Matthews, and Misti Shaw. 2013. "Music Discovery Requirements: A Guide to Optimizing Interfaces." *Notes* 69 (3): 494–524.

Onttonen, Esa. 2017. "The Arranger: Creating a Tool for Real-Time Orchestration and Notation on Mobile Devices." Master's thesis, Aalto University.

Peters, Charles. 2017. "Acquiring New Music from Unconventional Sources: PDF Copies in the Library." Paper presented at the Congress of the International Association of Music Libraries, Archives and Documentation Centres, Riga, Latvia, June, 18–22, 2017. http://www.iaml.info/sites/default/files/pdf/peters_pdf_copies_iaml_2017_presentation_notes_with_slides.pdf.

Rebelo, Ana, Ichiro Fujinaga, Filipe Paszkiewicz, Andre R. S. Marcal, Carlos Guedes, and Jamie S. Cardoso. 2012. "Optical Music Recognition: State-of-the-Art and Open Issues." *International Journal of Multimedia Information Retrieval* 1 (3): 173–90.

Rivera, Edward. 2017. *Book Publishing in the US.* IBISWorld Industry Report 51113. November 2017. http://clients1.ibisworld.com/reports/us/industry/default.aspx?entid=1233.

Scott, Darwin F. 2015. "Digital Scores in Music Libraries: The Status Quo and a New Response from MLA." Paper presented at the Annual Meeting of the Music Library Association, Denver, Colorado, February 25—March 1, 2015. https://vimeo.com/121103009.

Scott, Rachel E. 2013. "The Edition-Literate Singer: Edition Selection as an Information Literacy Competency." *Music Reference Services Quarterly* 16 (3): 131–40. https://doi.org/10.1080/10588167.2013.808941.

Serra, Xavier. 2017. "The Computational Study of a Musical Culture through Its Digital Traces." *Acta Musicologica* 89 (1): 24–44.

Steyn, Jacques, ed. 2013. *Structuring Music through Markup Languages: Designs and Architectures.* Hershey, PA: IGI Global.

Sturm, George. 2000. "Music Publishing." *Notes* 56 (3): 628–34.

Szeto, Kimmy. 2018. "Keeping Score, Digitally." *Music Reference Services Quarterly* 21 (2): 1–4. https://doi.org/10.1080/10588167.2018.1455027.

Tanaka, Atau. 2000. "Musical Performance Practice on Sensor-Based Instruments." *Trends in Gestural Control of Music* 13:389–405.

Underwood, Kent. 2016. "Scores, Libraries, and Web-Based, Self-Publishing Composers." *Notes* 73 (2): 205–40. https://doi.org/10.1353/not.2016.0115.

Underwood, Kent, and Robin Preiss, curators. n.d. New York University Collection of Contemporary Composers' Websites (online database). Accessed April 22, 2018. http://www.archive-it.org/collections/4049.

Vromman, Laurent. n.d. "Guitar Pro 4.06 File Format Description." Accessed April 22, 2018, http://dguitar.sourceforge.net/GP4format.html.

Walters, William H. 2013. "E-Books in Academic Libraries: Challenges for Acquisition and Collection Management." *portal: Libraries and the Academy* 13 (2): 187–211.

———. 2014. "E-books in Academic Libraries: Challenges for Sharing and Use." *Journal of Librarianship and Information Science* 46 (2): 85–95. https://doi.org/10.1177/0961000612470279.

Winget, Megan A. 2008. "Annotations on Musical Scores by Performing Musicians: Collaborative Models, Interactive Methods, and Music Digital Library Tool Development." *Journal of the American Society for Information Science and Technology* 59 (12): 1878–97. https://doi.org/10.1002/asi.20876.

Kimmy Szeto is an assistant professor and metadata librarian at Baruch College, City University of New York, where he oversees metadata management for digital resources. His recent research on the technical and conceptual tensions between cataloging practice and the linked data environment plays a pivotal role in shaping the modeling of linked data standards in the library community. He is an active presenter at the Music Library Association and the International Association of Music Libraries, and has published in *Notes*, the *Journal of Web Librarianship*, the *Journal of Electronic Resources Librarianship*, and the *Encyclopedia of Information Science and Technology*. Outside the library and academia, Kimmy can be heard as a chamber arranger of symphonic works and as a collaborative pianist in concert halls and theaters around New York City.

Paper 6

Ontology for Voice, Instruments, and Ensembles (OnVIE): revisiting the medium of performance concept for enhanced discoverability

Code4Lib Journal, 54

https://journal.code4lib.org/articles/16608 code{4}lib JOURNAL

Mission Editorial Committee Process and Structure Code4Lib

Search

Issue 54, 2022-08-29

Ontology for Voice, Instruments, and Ensembles (OnVIE): Revisiting the Medium of Performance Concept for Enhanced Discoverability

Medium of performance—instruments, voices, and devices—is a frequent starting point in library users' search for music resources. However, content and encoding standards for library cataloging have not been developed in a way that enables clear and consistent recording of medium of performance information. Consequently, unless specially configured, library discovery systems do not display medium of performance or provide this access point. Despite efforts to address this issue in the past decade in RDA, MARC, and the linked data environment, medium of performance information continues to be imprecise, dispersed across multiple fields or properties, and implied in other data elements. This article proposes revised definitions for "part," "medium," "performer," and "ensemble," along with a linked data model, the Ontology for Voice, Instruments, and Ensembles (OnVIE), that captures precise and complete medium of performance data reflecting music compositional practices, performance practices, and publishing conventions. The result is an independent medium of performance framework for recording searchable and machine-actionable metadata that can be hooked on to established library metadata ontologies and is widely applicable to printed and recorded classical, popular, jazz, and folk music. The clarity, simplicity, and extensibility of this model enable machine parsing so that the data can be searched, filtered, sorted, and displayed in multiple, creative ways.

By Kimmy Szeto

Introduction

Medium of performance—the instruments, voices, and devices at the library users' disposal—is a frequent starting point in their search for music resources, and an identifying element of musical works and expressions (Ostrove, 2001). However, the limited availability and functionality of medium of performance as an access point has persisted from the physical card catalog, through the MARC era, into current developments of linked data ontologies.

Uniquely important to music resources, medium of performance refers to the tools involved in expressing a musical work. Historically, the term "medium" is not well defined in cataloging rules, and some instructions and practices are incompatible with linked data modeling practice today. Medium of performance has been recorded as supplemental information in the title area (Coyle, 2011) and in the subject area (Subject Analysis Committee, 2017), and did not receive much attention as an independent data element until the release of the current generation of library cataloging standard Resource Description and Access (RDA) in 2010. Inheriting the basic outline from earlier standards, RDA instructs catalogers to record a list of instruments, voices, and ensembles, followed by the number of parts and total number, and, when necessary, further identifying characteristics.

Cataloging practice continues to treat medium of performance as a single entity. The reality, however, is not as simple. Tracing a musical work from its composition to performance, subsumed in what is broadly termed "medium of performance" is a network of relationships involving the composer (musical parts), the publisher (published scores and parts), the instruments/voices (mediums), the performers, and the ensembles. From the library users' standpoint, the single list of instruments/voices and numbers in the data does not always match the score and parts in the actual publication, and often does not provide enough information about the specific instruments, devices, and players required for planning a reading or a performance.

While these more complex relationships pose a challenge for the MARC format to encode, linked data models offer the capability. Linked data —a set of technologies and practices that foster publishing and connecting structured data on the web— have been increasingly embraced in the past decade by the library community as a means to provide open access to its richly curated bibliographic catalogs. Although some attempts have been made to refine the medium of performance element in linked data models, such as in the Performed Music Ontology (PMO), so far, no library linked data model has been built out beyond alignment with RDA/MARC format and conversion of existing data.

In the context of expanding the conception of the medium of performance element, this article lays out broadened definitions for "part," "performer" and "ensemble" as described in Szeto (2017), proposes a revised definition for the term "medium," and, based on these revised definitions, presents a new data model, the Ontology for Voices, Instruments, and Ensembles (OnVIE). OnVIE gives medium of performance additional dimensions by simultaneously capturing and relating the composer's intended performing forces, published parts, specific instruments, voices, and devices required/used, as well as specific performers. The clarity, simplicity, and extensibility of this model enable more nuanced machine parsing so that the data can be searched and displayed in multiple, creative ways.

Modeling Medium of Performance

This paper will first discuss the development of OnVIE in the context of RDA, the current content standard, as encoded in MARC field 382. The next section will turn to the Performed Music Ontology, an extension for music resources specifically developed as an extension for BIBFRAME, the general linked data model being developed for library bibliographic data to replace MARC.

Recent Cataloging Practice

Prior to the 2000s, the tripartite division of author, title, and subject of the card catalog resulted in cataloging rules and decades of practices that embedded medium of performance information in the uniform title when it is necessary to distinguish between identical titles and in subject heading form subdivisions (Elmer, 1960). The MARC format did offer field 048 for medium of performance, but the fields held only codes and numbers for voices, instruments, and ensembles, which were not readily decipherable to catalogers and users alike. This coded field fell into disuse, and users mainly relied on a free text note and clues from the title and subject headings. The dispersal of structured data complicated display and indexing, resulting in a limited ability for users to search directly or filter medium of performance search results (Subject Analysis Committee, 2017). This is particularly problematic for compilations, vocal music, folk music, jazz, and recorded popular music where medium of performance is generally not explicitly stated bibliographically (Newcomer et al., 2013, section II.D).

These limitations led to coordinated efforts to raise the visibility of this data element. In 2007, the Library of Congress, collaborating with a range of stakeholders, initiated developments in three areas: a dedicated faceted vocabulary Library of Congress Medium of Performance Thesaurus (LCMPT) was launched in 2014 (Library of Congress, 2014), a new MARC field 382 for an expanded encoding of medium of performance data based on RDA instructions was established in 2010 (Library of Congress, 2020), and programmatic changes that extracted existing authority data from the uniform title to be placed into the 382 field took place on WorldCat (Library of Congress, 2012). With these three areas in place, programmatic retrospective implementation of faceted vocabulary in both authority and bibliographic records continues to this day (Mullin, 2018; Subject Analysis Committee, 2022).

Below is an excerpt from a MARC record on WorldCat (OCLC number 989164116) that illustrates how medium of performance information is recorded in the uniform title (field 240) and the subject headings (field 650) and coded in fields 048 and 382. It might not be immediately apparent, from the way the uniform title and subject headings are constructed and structured, that this record is based on a musical score for solo harpsichord with piano accompaniment. (There is no string orchestra!) What might (or might not!) help library users is the note in field 500, a phrase typically used for this situation in music cataloging practice. Unfortunately, the note is not present in the actual WorldCat record, but was added here by the author to illustrate the difficulties posed to human readers and computer algorithms alike. The issue with clarity of semantics was identified and critiqued in Coyle (2011), and the retrospective implementation of LCMPT and field 382 to address this issue was detailed in Mullin (2018).

```
048 ##  $b kc01 $a ka01
240 10  $a Concertos, $m harpsichord, string orchestra, $n BWV 1052, $r D
382 01  $b harpsichord $n 1 $a piano $n 1 $s 2 $2 lcmpt
500 ##  $a Acc. arr. for piano.
650 #0  $a Concertos (Harpsichord with string orchestra) $v Solo with pia
```

Data Modeling Considerations

While these efforts have enabled structured medium of performance data in bibliographic records using LCMPT as the controlled vocabulary, the implementation remains less than ideal. RDA inherited the broad outlines from earlier standards—"the instrument, instruments, voice, voices, etc. for which a musical work was originally conceived" (RDA 6.15.1.1), and then "record each instrument…" (RDA 6.15.1.4) or "record an/the appropriate term…" that groups instruments and voices by family or into an ensemble (RDA 6.15.1.6 to 6.15.1.10). These terms, depending on their characteristics, are encoded in MARC 382 subfields $a, $d, $d, and $p. After each term, RDA instructs the cataloger to record the number of parts (RDA 6.15.1.3). This number is then encoded in MARC 382 subfields $e, $n or $r, which are defined for the number of performers or ensembles. In addition, MARC 382 offers subfields $s and $t for the total numbers of performers and ensembles.

This practice creates some data subfields that are not unambiguous in some cases and not atomic in some others. In the course of resolving ambiguity and atomicity issues, two fundamental principles of linked data design for the semantic web, it became apparent that separate definitions for part, medium, performer, and ensemble could offer the overarching solution.

Ambiguity and Atomicity Issues

The RDA instruction to record the number of "parts" which is then encoded in the "number of performers" subfield $n in MARC field 382 creates a semantic ambiguity. The OnVIE model will treat part and performer as two separate concepts. Another conflation occurs in the usage of terms such as "percussion" and "continuo" as mediums, when these mediums are actually parts referring to a group of instruments. The OnVIE model will provide clarity by allowing the individual instruments to be linked to these parts.

Another ambiguous practice is conflating individual instruments/voices with ensembles, and encoding them in the same MARC subfields, even though ensembles are not themselves instruments or voices. However, it is not always clear what exactly forms an ensemble. Is the string quartet an ensemble of instruments? Is it a group of performers? Or is it a group of parts? In the OnVIE model, individual parts are used as the starting point, and a group of parts will form an ensemble. This way, all of the ensemble's constituent parts will always be known, as will the mediums be linked to each part.

Separate instructions are given in RDA for music intended for one performer to a part as opposed to more than one performer to a part. The attributes of "solo" and "accompaniment" are also treated separately. However, these are not characteristics of the mediums, but are characteristics of the parts. (For example, the instrument violin itself cannot possess the quality of "solo" but a violin part can.) In OnVIE, the number of performers to a part, the solo status, and the accompaniment status will all be treated as refinements to the part, rather than as separate classes of entities.

More specifically, RDA instructs the cataloger to omit the accompanying keyboard instrument (such as a piano) in a classical song for solo voice (RDA 6.15.1.11). While the piano is not an unreasonable assumption among musicians, such an arbitrary exception causes the discovery system to return incomplete results, which could complicate a library user's search, especially for a vocalist looking for repertoire regardless of the accompaniment.

Special Issue: Number of Hands

RDA instructs the "number of hands," if other than two, to be recorded (RDA 6.15.1.5.1). This poses an impossibility in the encoding mechanism in MARC field 382. While the number of performers is encoded in subfield $n, there is nowhere to encode the number of instruments, or which hand is being used. While "piano, 4 hands," where two pianists use all four hands on a single piano, is a fairly common genre, piano duets—two pianos and two pianists—is not uncommon either, especially in more recent repertoire with the two pianos tuned a quarter tone apart (for example, *Chiaroscuro for two pianos (one tuned down a 1/4 tone)* by John Corigliano (published 2011)). Yet these are encoded identically as "$a piano $n 2," and library users will need to inspect the free text notes, the subject heading, or the title. Suppose a piece of music was composed for two pianos with two players, each using only the left hand, or, suppose another piece of music was composed for two pianos with two players, one using both hands, the other using only the left hand, the result is still the same: "$a piano $n 2." Figure 1, adapted from Szeto (2016, slide 22), further illustrates this issue with combinations of two pianos and six hands. Meanwhile, a piece of music composed for two pianos and one player (for example, *Trois hommages for 2 pianos (tuned a quarter tone apart) 2 hands* by Georg Friedrich Haas (published 2009)) would be encoded as "$a piano $n 1," identical to solo piano music (one piano, one player).

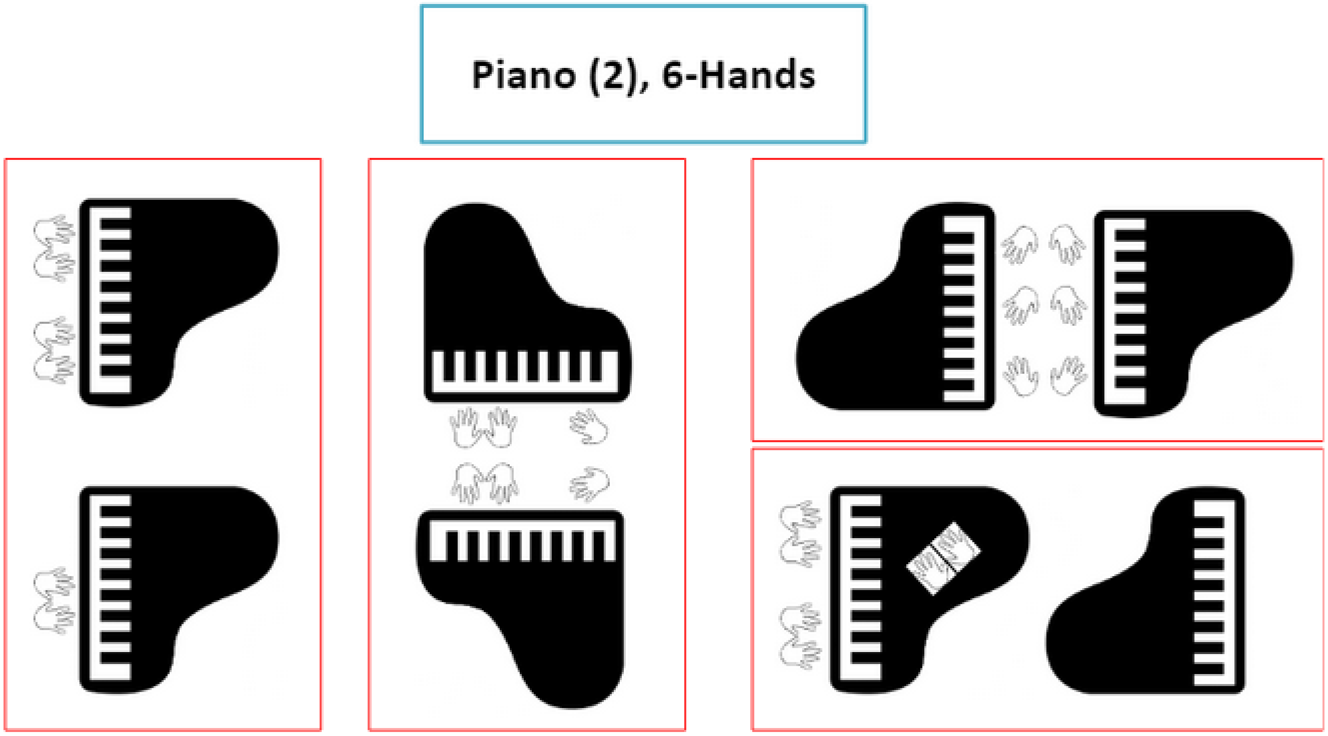


Figure 1. Combinations of pianos and hands that cannot be differentiated in RDA descriptions or encoded in MARC field 382.

At present, when these distinctions, or any other distinctions discussed above, cannot be expressed in the MARC 382 field, the cataloger is instructed to provide clarification in a free text note in subfield $v. In the OnVIE model, separate treatments for part and performer will remove this issue.

Additional RDA Medium of Performance Characteristics

RDA includes instructions for three more attributes: pitch and range of instruments, doubling, and alternative.

Pitch and Range

The "pitch" of the instrument here refers to musical instruments that are "transposed," where playing with the same set of fingering and sound production techniques on the instrument yields a different set of pitches (for example, clarinet in B-flat). What RDA referred to as "range" refers to instruments that come in common sizes (for example, the saxophone family includes soprano, alto, tenor, and baritone). Pitch and range are recorded only when the cataloger considers them important for identification and access (RDA 6.15.1.5.1) and appears in MARC field 382 as a free text note in subfield $v. The OnVIE model will include further refined properties for these, applied separately to parts and mediums.

Doubling

Instrumental doubling refers to a musical part that instructs a single performer to use more than one instrument. When recorded, the medium is encoded in MARC 382 subfield $d. This structure seems straightforward for a traditional orchestral doubling, such as a flute player who also plays the piccolo. This is not so straightforward for other parts that are less clearly defined, for example, an orchestral percussion part where multiple players share a set of percussion instruments, or where a timpani player occasionally joins other percussionists (timpani is recorded separately from percussion). Currently, the instruction is to record the term "percussion" alone. Neither RDA nor MARC offers a solution for this entanglement between performers, mediums, and parts. Separating them in the OnVIE model will remove this issue.

Alternative

RDA instructions for "alternative instruments" (RDA 6.15.1.5.3) invite further refinement. In current practice, "alternative" is used for a range of situations where one or more mediums deviate from the original. While some composers expressly indicate that the identical part may be used on more than one instrument (for example, the "Flute or Violin" part in the jazz composition *Out of the Cool* by David Heath (as published in 1986)), in most cases, what are also considered alternatives range from slightly different (for example, clarinet music transcribed for the viola; see Swanson, 2003, pp. 13-15 for the process of compiling the repertoire list which could have been aided by a direct medium of performance search), to quite different (for example, a continuo part realized for piano), to drastically different (for example, orchestra music in operatic works arranged for piano with no change in the vocal parts). The OnVIE model will be able to precisely indicate the first case where the part is identical, and a hook is provided for future extensions that capture the various degrees of change. There are also situations where the alternative involves a role change in one or more parts (for example, the arrangements of the *Scherzo* movement in Robert Schumann's second symphony (as published in Szeto, 2010a and 2010b) where the first violin part could be a solo part or an ensemble part; the arrangement of the overture to Leoš Janácek's *From the House of the Dead* (as published in Szeto, 2013), where the violin part changed to solo and the trumpet part could optionally be offstage). Currently, data cannot be simultaneously encoded as alternative (in $p) and as solo (in $b) in MARC field 382, whereas OnVIE imposes no restrictions on these refinements. Equally important is the consideration of whether the alternative is so different that the music should be considered a separate musical expression or an entirely new musical work. This is outside the scope of this article but was discussed in more detail in Szeto (2017), and the data modeling is an active area of investigation under various keywords such as music information retrieval, arrangement, versions, and annotation (Weiß et al., 2021; Lewis et al., 2022).

More specifically, the Library of Congress Policy Statements for RDA 6.18.1.4 instructs catalogers whether or not to consider alternatives or arrangements in a range of conditions, such as slight alterations to instrumentation in an orchestral piece, early music composed before 1800 performed on modern instruments, and a change from vocal to instrumental in the popular music idiom. While such practices to make an either/or judgment are essential for the classification and collocation of physical materials, the OnVIE model will remove much of these artificial criteria by simply allowing all parts and mediums to be recorded completely and precisely.

Another common situation is alternative voices (such as a song for soprano sung by a tenor, or for alto sung by a bass). This is an ambiguity issue where the voice part is the same but the property of the medium differs. In the OnVIE model, soprano/alto/tenor/bass can be indicated as refinements to the medium "voice."

BIBFRAME and the Performed Music Ontology

The Bibliographic Framework Initiative (BIBFRAME) is a Library of Congress initiative to develop a linked data alternative to MARC. The Library of Congress initially worked with the Music Library Association to model the medium of performance element but later abandoned the effort. Instead, the BIBFRAME model left "hooks" for a full model to be developed by a third party (Szeto et al., 2016, p. 32).

In response, the Performed Music Ontology (PMO, 2021) was developed as an extension of the BIBFRAME ontology with a focus on describing performed music. The medium of performance portion of the PMO model tracks closely with RDA and MARC practices. PMO begins with making the distinction between the declared medium, which is stated by the composer or in a reference source, and the performed medium, which was used in the actual performance and not necessarily the same as the declared medium. The PMO has separate properties for individual and ensemble mediums, a separate class for "part," as well as properties to connect performers to their instruments, voices, and dramatic roles. However, PMO follows RDA instructions where solo, alternative, and doubling are separately addressed, while other attributes are recorded as literals (free text) in a single catch-all "part type" property. The OnVIE model will break away from the RDA/MARC structure, and will provide an extensible structure for characteristics to be encoded as linked data with clear semantics.

Unlike PMO where medium of performance classes depend on linking to the parent ontology at multiple points (*bf:NotatedMusic* for *pmo:DeclaredMedium*, *bf:Audio* for *pmo:PerformedMedium*, *bf:Contribution* for the *pmo:IndividualMOP* and *pmo:MusicPart*), the OnVIE model exists in a self-contained, independent space requiring only a single "hook" to the musical resource being described. The first contact on this hook is "Parts." By beginning with this layer, the OnVIE model no longer requires differentiation between mediums that were notional or actual.

A New Data Model for Medium of Performance

Ontology for Voice, Instruments, and Ensembles

The primary motivation, and innovation, of the OnVIE model is capturing the relationship between part, medium, performer, and ensemble, which form the four main classes of the ontology. The first three main classes follow a loop: Part connects to Medium, Medium connects to Performer, and Performer loops back to Part. The class Ensemble forms a branch by grouping one or more Part entities. These relationships are shown in Figure 2.

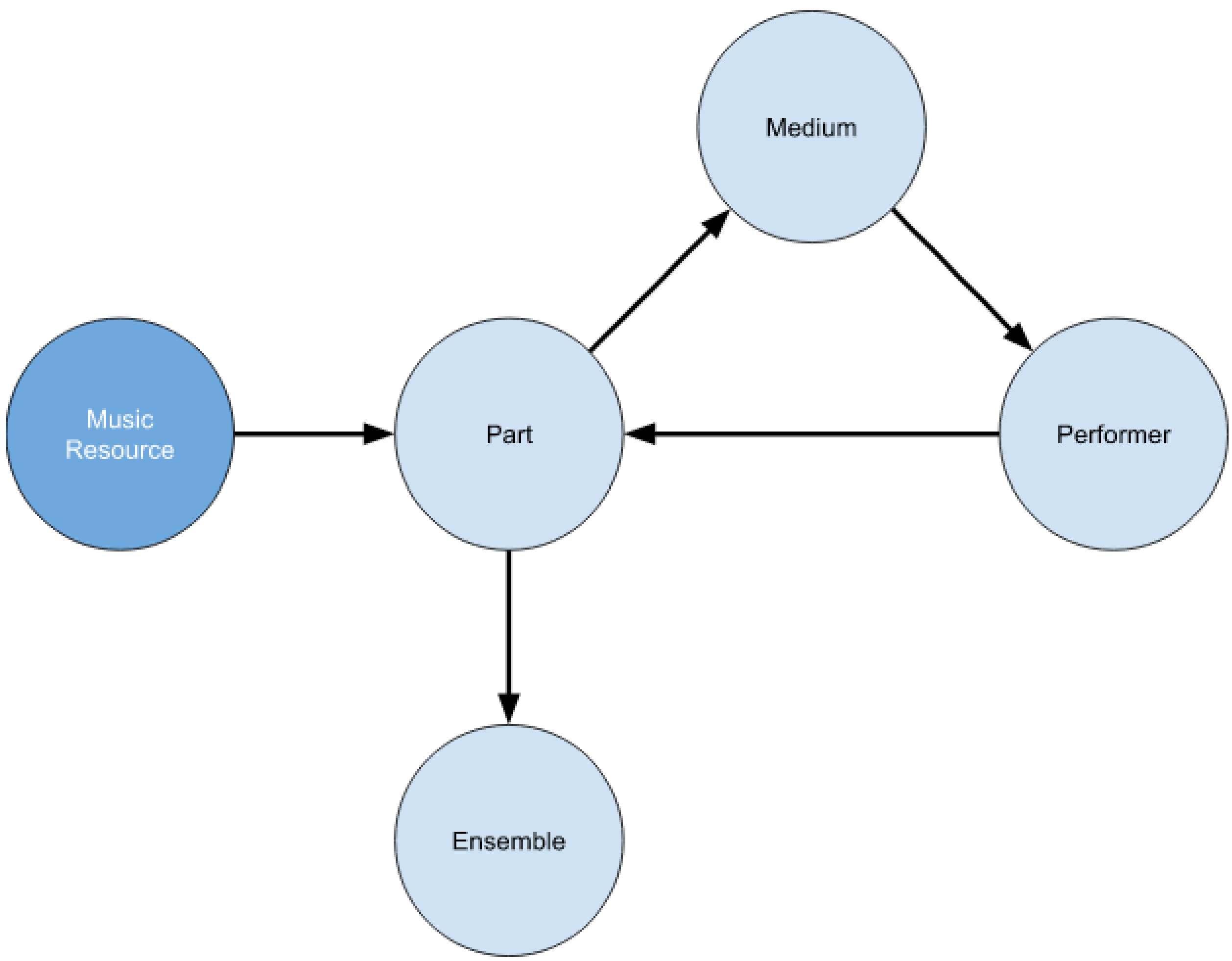


Figure 2. Class relationships in the Ontology for Voice, Instruments, and Ensembles.

All further refinements to the model fall within this framework. The list of refinements presented here consists of data that are currently recorded in cataloging practice, fixes to known issues, with many more properties that cover characteristics of specifical musical settings, parts, instruments, voices, devices, etc. OnVIE's simple framework allows it to be readily extensible through additional refinements.

Definitions

Part

Class: MusicPart
Scope: a series of musical events, abstracted from a musical work/expression, generally independent from other such abstractions from the same musical work/expression, which holds a consistent association with one or more mediums and/or performers
Note: A part is not equivalent to, but may be related to, a "printed part" (sheet music for an individual voice/instrument that does not contain notation for other voices/instruments of an ensemble), or a "voice" or "line" (individual melodies of polyphonic musical composition)
Examples: "Violin" ; "Percussion" ; "Mezzo-Soprano" ; "Horns"

Medium

Class: MusicMedium
Scope: a tool of sound production for a musical work/expression
Note: Mediums include human voices, musical instruments, devices, and other means
Examples: "violin" ; "marimba" ; "voice" ; "saxophone" ; terms in LCMPT and UNIMARC codes for individual voice, instruments, and devices

Performer

Class: MusicPerformer
Scope: an agent responsible for expressing or actuating one or more parts
Note: Performers include humans, machines, computers, and other entities
Examples: "Violinist" ; "Percussionist" ; "Vocalist" ; "Saxophonist" ; entities referring to actual persons in vocabularies such as LCNAF and VIAF

Ensemble

Class: MusicEnsemble
Scope: a group of parts in a musical work/expression
Note: in this model, only part groupings are used, even for ensembles traditionally known for the personas or the instruments (for example, string quartet)
Examples: "Orchestra" ; "Percussion quartet" ; "Children's choir" ; "Jazz combo" ; ensemble terms in LCMPT and UNIMARC codes ; LCNAF, VIAF, or other vocabularies for names

Classes and Properties

The OnVIE model is built with the library users' starting point in mind: the instruments, voices, and devices (class MusicMedium). The connecting layer between the medium and the music resource is the concept of "part" (class MusicPart). The term "part" has been occasionally used in definitions and instructions, and occasionally conflated with voices, instruments, and performers, as discussed above. Here, a new definition has been developed for this layer, distinct from MusicMedium and MusicPerformer.

The OnVIE model begins with at least one MusicPart. A MusicPart can link to other MusicPart entities when alternatives or further subdivisions are present, or can link to printed published parts which could differ. Every MusicPart is required to have at least one MusicMedium, even if it is unmediated (such as a spectator). MusicPart can link to more than one entity in MusicMedium, such as in the case of instrumental doubling or a single percussion part calling for multiple instruments. For musical scores, it ends here. For performed music, including popular, folk, and jazz that exist only as sound recordings, performers (class MusicPerformer) are linked from MusicMedium, and each MusicPerformer loops back to one or more entities in MusicPart.

Ensembles are not considered mediums, but are a separate class (class MusicEnsemble), formed by an aggregate of individual parts. This departure from current cataloging practice prevents ensemble terms from being used as mediums, but rather encourages a complete accounting of mediums involved.

The four main properties of this model connect these classes. Refinements are provided to describe each in further detail. The list of refinement properties presented here was drawn from the current models as well as from the author's own experience as an ensemble librarian, and is by no means exhaustive.

What is not included in this model are the numbers. Rather than requiring catalogers and metadata creators to supply the number of parts, performers, and ensembles, the granularity of the model enables linked data interpreter software to perform the counting. This method not only provides the flexibility to produce separate counts for parts, mediums, performers, and ensembles, it also removes the uncertainty and detailed analysis (or guesswork!) required to arrive at the number of performers needed for group-oriented parts such as percussion and continuo. By interpreting the refinements, machine counting can also provide the total, as well as further numerical breakdowns for soloists, accompanying performers, ensembles, and voice/vocal parts.

The Ontology

Table 1. Classes.

Class	Subclass of
MediumOfPerformance	Musical Works/Expressions
MusicMedium	MediumOfPerformance
MusicPart	MediumOfPerformance
MusicPerformer	MediumOfPerformance
MusicEnsemble	MediumOfPerformance

Table 2. Properties.

Property	Use with	Expected value
hasMusicPart	MediumOfPerformance ; MusicPart	MusicPart
isMusicPartOf	MusicPart	MusicPart ; MediumOfPerformance
hasMusicMedium	MusicPart	MusicMedium
isMusicMediumOf	MusicMedium	MusicPart
hasMusicPerformer	MusicMedium	MusicPerformer
isMusicPerformerOf	MusicPerformer	MusicMedium
isResponsibleForMusicPart	MusicPerformer	MusicPart
isPerformedBy	MusicPart	MusicPerformer

Table 3. Refinements.

Property	Use with	Expected value	Notes	Examples
rdfs:label	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	literal		
xml:lang	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	URI	Language code of the term	
source	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	URI	Source of information	
sourceType	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	“transcribed” ; “recorded” ; “published” ; “inferred” ; “editorial” ; “programmatic update” ; etc.		
sourceNote	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	literal		“First page of music”
alternative	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	“Is alternative” ; “Is not alternative” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
alternativeType	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	URI	* hook for the full consideration of types of musical alternation/arrangement/transcription/adaptation	
alternativeNote	MediumOfPerformance ; MusicPart ; MusicMedium ; MusicEnsemble	literal		“identical” ; “arranged for viola” ; “transcribed for solo piano” [concerto, jazz] ; “piano reduction” [orchestral accompaniment] ; “adapted for the violin” [folk music] ; “combined percussion part for one player”
partNumber	MusicPart	whole number >=0		“Violin 1” ; “Percussion 2”
playerToAPart	MusicPart	“Specified” ; “Multiple” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
playerToAPartNumber	MusicPart	whole number >=0		
solo	MusicPart ; MusicMedium ; MusicEnsemble	“Is solo” ; “Is not solo” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
accompaniment	MusicPart ; MusicMedium ; MusicEnsemble	“Is an accompaniment” ;		

Property	Use with	Expected value	Notes	Examples
		“Is not an accompaniment”		
optional	MusicPart ; MusicMedium	“Is optional” ; “Is not optional” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
ad lib	MusicPart ; MusicMedium	“Is ad lib” ; “Is not ad lib” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
offstage	MusicPart ; MusicMedium	“Is offstage” ; “Is not offstage” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
obligato	MusicPart ; MusicMedium	“Is an obligato part” ; “Is not an obligato part” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
amplified	MusicPart ; MusicMedium	“Is amplified” ; “Is not amplified” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
prerecorded	MusicPart ; MusicMedium	“Is prerecorded” ; “Is not prerecorded” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
periodInstrument	MusicPart ; MusicMedium	“Is a period instrument” ; “Is not a period instrument” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
periodInstrumentNote	MusicPart ; MusicMedium	URI or literal		“Baroque” [flute]
fingeringSystem	MusicPart ; MusicMedium	URI		“German” [recorder]
tuningSystem	MusicPart ; MusicMedium	URI		“Just intonation” ; “pythagorean” ; “equal temperament”
tuningReferencePitch	MusicPart ; MusicMedium	URI	The pitch name of the tuning reference pitch.	“A”
tuningReferenceFrequencyHz	MusicPart ; MusicMedium	number >=0	The frequency of the tuning reference pitch in hertz.	“432”
scordatura	MusicPart ; MusicMedium	“Is tuned scordatura” ; “Is not tuned scordatura” ; “Performer’s choice” ; “Unspecified” ; “Unknown”	Tuning of a western string instrument which deviates from the standard tuning.	
tuningNote	MusicPart ; MusicMedium	literal		“Piano is tuned quarter tone flat” ; “Drop D” [guitar]
handsNumber	MusicPart ; MusicMedium	whole number >=1	Number of hands playing an instrument	
handsSide	MusicPart ; MusicMedium ; MusicPerformer	“Left” ; “Right” ; “Performer’s choice” ; “Unspecified” ; “Unknown”	Which hand is being used for playing an instrument	
handsNote	MusicPart ; MusicMedium ; MusicPerformer	literal		“piano (2), 3 hands”
doubleBassCExtension	MusicPart ; MusicMedium	“Requires a C Extention” ; “Does not requires a C Extension” ; “Performer’s choice”		

Property	Use with	Expected value	Notes	Examples
		; “Unspecified” ; “Unknown”		
fluteBFoot	MusicPart ; MusicMedium	“Requires a B Foot” ; “Does not require a B Foot” ; “Performer’s choice” ; “Unspecified” ; “Unknown”		
instrumentMute	MusicPart ; MusicMedium	URI		
instrumentMuteNote	MusicPart ; MusicMedium	literal		
instrumentDimension	MusicPart ; MusicMedium	URI	* hook for measurements of musical instruments	
instrumentDimensionNote	MusicPart ; MusicMedium	literal		“26 inch” [timpani]
instrumentSize	MusicPart ; MusicMedium	URI	* hook for instrument sizes	
instrumentSizeNote	MusicPart ; MusicMedium	literal		“Three-quarter” [guitar] ; “Concert” [ukelele]
instrumentPitch	MusicPart ; MusicMedium	URI	Pitch of single-pitched musical instruments	“C4” [crotale]
instrumentTransposition	MusicPart ; MusicMedium	URI	Key of transposing instruments	“A” [clarinet]
instrumentTranspositionNote	MusicPart ; MusicMedium	literal		“Clarinet in A”
instrumentRangeNumber	MusicPart ; MusicMedium	whole number >=0	Pitch range of a musical instrument in number of half steps	“60” [marimba]
instrumentRangeLowest	MusicPart ; MusicMedium	URI	Lowest pitch of a musical instrument	“C2” [marimba]
instrumentRangeHighest	MusicPart ; MusicMedium	URI	Highest pitch of a musical instrument	“C7” [marimba]
instrumentNote	MusicPart ; MusicMedium	literal		“5 octaves” [marimba]
voiceType	MusicPart ; MusicMedium	URI		“Mezzo soprano” ; “Contralto”
voiceWeight	MusicPart ; MusicMedium	URI		“Spinto” ; “Soubrette”
voiceTessitura	MusicPart ; MusicMedium	URI		“High” ; “Medium” ; “Low”
voicePitchLowest	MusicPart ; MusicMedium	URI	Lowest pitch required of the vocalist	
voicePitchHighest	MusicPart ; MusicMedium	URI	Highest pitch required of the vocalist	
voiceNote	MusicPart ; MusicMedium	literal		
technicalRequirement	MusicPart ; MusicMedium	URI	* hook for computer / recording carrier / playback device information	
technicalRequirementNote	MusicPart ; MusicMedium	literal		“Requires an 8 track player”

Next Steps

With the OnVIE model laid out conceptually in this paper, the next steps would be to formalize the ontology with modeler software, and, with the help of the library community, integrate into library linked data editors and test a range of use cases to further fine tune semantics, data constraints, and documentation. In the course testing, standardized vocabularies can be developed for the many characterizations in the list of refinements, possibly in alignment with open platforms such as Wikidata. Some suggestions for these “hooks” can be found in the “Notes” column in Table 3. Finally, pathways to publishing and maintaining the ontology can be explored.

Further analysis can be made in relation to the UNIMARC encoding standard, as well as to library-adjacent ontologies such as DoReMus, developed for analysis and visualization of music data, the Music Ontology, which focuses on capturing production of musical events, and MusicBrainz, which is widely used for sound recordings. As none of these ontologies currently includes a model built out for medium of performance, developing a mechanism to hook OnVIE on to them would be a worthwhile investigation. Another potentially fruitful area of study would be to align MusicParts in OnVIE with other ontologies where the concept is also used, such as the “Observations” object in the “Musicological Objects” layer in Lewis et al. (2022).

Conclusion

Although linked data are designed to be machine-actionable, it is humans who ultimately employ the mediums with their voices, instruments, and other tools, to create music. It is also humans who are ultimately responsible for expressing and actuating each part of a musical work. This new Ontology for Voices, Instruments, and Ensembles, when hooked on to linked data bibliographic systems, will enable a medium of performance access point at a fine level of precision and completeness. Library users will be provided a more straightforward path not only toward identifying and selecting music resources, but also toward discovering additional insights into the evolution of performing forces in the history of music making, a whole new area of humanistic studies previously hidden in plain sight.

Bibliography

Corigliano, J. (2011). Chiaroscuro: For two pianos (one tuned down a 1/4 tone) (ED 4435, HL 50490191). G. Schirmer.

Coyle, K. (2011). MARC as metadata: A start. Code4lib Journal, 14. https://journal.code4lib.org/articles/5468.

Elmer, M. (1960). The music catalog as a reference tool. Library Trends, 8(4), 529-538. https://www.ideals.illinois.edu/bitstream/handle/2142/5907/librarytrendsv8i4f_opt.pdf.

Haas, G. F. (2009). Trois hommages: Für zwei Klaviere (im Vierteltonabstand gestimmt) zu zwei Händen (1982/84) (UE 34 693). Universal Edition.

Heath, D. (1986). Out of the cool: For flute (or violin) and piano (CH55693). Chester Music.

Lewis, D., Shibata, E., Saccomano, M., Rosendahl, L., Kepper, J., Hankinson, A., Siegert, C., and Page, K. (2022). A model for annotating musical versions and arrangements across multiple documents and media. In 9th International Conference on Digital Libraries for Musicology (DLfM2022) (pp. 10-18). Association for Computing Machinery. https://doi.org/10.1145/3543882.3543891.

Library of Congress. (2012). Summary of programmatic changes to the LC/NACO Authority File: What LC-PCC RDA catalogers need to know. https://www.loc.gov/aba/rda/pdf/lcnaf_rdaphase.pdf.

Library of Congress. (2014). Library of Congress launches Medium of Performance Thesaurus for music. https://www.loc.gov/catdir/cpso/medprf-list-launch.html.

Library of Congress. (2020). Content designator history. In 382 – Medium of Performance. https://www.loc.gov/marc/bibliographic/bd382.html.

Mullin, C. (2018). An amicable divorce: Programmatic derivation of faceted data from Library of Congress Subject Headings for music. Cataloging & Classification Quarterly, 56(7), 607-627. https://doi.org/10.1080/01639374.2018.1516709.

Newcomer, N. L., Belford, R., Kulczak, D., Szeto, K., Matthews, J., & Shaw, M. (2013). Music discovery requirements: A guide to optimizing interfaces. Notes, 69(3), 494-524. http://dx.doi.org/10.1353/not.2013.0017.

Online Computer Library Center. (2022, May 24). OCLC 989164116. A publicly accessible version of this record as imported by University of California, Berkeley can be retrieved at https://search.library.berkeley.edu/discovery/sourceRecord?vid=01UCS_BER:UCB&docId=alma991033194509706532&recordOwner=01UCS_NETWORK.

Ostrove, G. E. (2001). Music subject cataloging and form/genre implementation at the Library of Congress. Cataloging & Classification Quarterly, 32(2), 91-106. https://doi.org/10.1300/J104v32n02_08.

Performed Music Ontology: Documentation (2021, May 6). https://github.com/LD4P/PerformedMusicOntology/tree/main/Documentation.

RDA Toolkit. (2022). American Library Association. https://access.rdatoolkit.org/.

Subject Analysis Committee. (2017). A Brave New (Faceted) World: Towards Full Implementation of Library of Congress Faceted Vocabularies. Association for Library Collections & Technical Services. http://hdl.handle.net/11213/8146.

Subject Analysis Committee. (2022). Retrospective Implementation of Library of Congress Faceted Vocabularies. Association for Library Collections & Technical Services. http://hdl.handle.net/11213/17998.

Swanson, C. M. (2003). Adding to the viola repertoire by arranging: A study on methods of arranging music for viola from clarinet, with an original arrangement of the Saint-Saens Clarinet Sonata in E-flat, Op. 167 (UMI number: 3107045) [Doctoral dissertation, University of Arizona]. ProQuest Information and Learning. http://hdl.handle.net/10150/280389.

Szeto K. (2010a). Concerto for violin and chamber ensemble arranged after Robert Schumann's Scherzo from symphony no. 2. Alexander Street Press Classical Scores Library, Volume IV. https://search.alexanderstreet.com/preview/work/bibliographic_entity%7Cscore%7C3643958.

Szeto K. (2010b). Robert Schumann's symphony no. 2 in C major arranged for chamber ensemble. Alexander Street Press Classical Scores Library, Volume IV. https://search.alexanderstreet.com/preview/work/bibliographic_entity%7Cscore%7C3643956.

Szeto K. (2013). Leoš Janácek: Ouvertüre zur Oper "Aus einem Totenhaus" für Solovioline und Ensemble (UE 36 249). Universal Edition.

Szeto, K. (2016, October 15). Untangling medium of performance for the linked data environment [presentation]. New York State-Ontario Chapter of the Music Library Association Fall Meeting, University of Toronto, Ontario, Canada. https://academicworks.cuny.edu/bb_pubs/1251/.

Szeto, K., Adams, A. D., Billet, K. E., Busselen, C., Kishimoto, K. S., LoPrete, A. A., McFall, L., Rondeau, S., Snyder, T. L., Soe Nyun, J. L., Vanden Dries, W. R., & Vermeij, H. (2016). Report of the CMC BIBFRAME Task Force to the Board of the Music Library Association. https://cmc.wp.musiclibraryassoc.org/wp-content/uploads/sites/5/2019/02/201602Task_Force_Report.pdf.

Szeto, K. (2017). The mystery of the Schubert Song. Notes, 74(1), 9-23. https://doi.org/10.1353/not.2017.0071.

Weiß, C., Zalkow, F., Arifi-Müller, V., Müller, M., Koops, H. V., Volk, A., & Grohganz, H. G. (2021). Schubert Winterreise Dataset: A Multimodal scenario for music analysis. Journal on Computing and Cultural Heritage, 14(2), 1-18. https://doi.org/10.1145/3429743.

About the Author

Kimmy Szeto is Associate Professor and Metadata Librarian at Baruch College, City University of New York, where he manages metadata for digital resources. His recent research focuses on the technical and conceptual tensions between cataloging practice and the linked data environment. Outside the

library and academia, Kimmy can be heard as a chamber arranger of symphonic works and as a collaborative pianist in concert halls and theaters around New York City.