Linked Data for Libraries, Archives and Museums

www.alastore.ala.org

Linked Data for Libraries, Archives and Museums

How to Clean, Link and Publish Your Metadata

Seth van Hooland and Ruben Verborgh

Neal-Schuman
An imprint of the American Library Association
Chicago | 2014
www.alastore.ala.org

First published in the United Kingdom by Facet Publishing, 2014.
This simultaneous U.S. Edition published by ALA Neal-Schuman,
an imprint of the American Library Association, 2014.

18 17 16 15 14   5 4 3 2 1

ISBN: 978-0-8389-1251-5 (paper)

www.alastore.ala.org

Contents

The authors

Seth van Hooland is an assistant professor at the Université libre de Bruxelles (ULB). After a career in the private sector for a digitization company, he obtained his PhD in information science at ULB in 2009. Following a post-doc position at the University Carlos III of Madrid, Seth joined the Information and Communication Science Department at ULB and became the academic responsible for their Master in Information Science. In the spring semester of 2014, he taught a special course on linked data at the Information School of the University of Washington. He is also active as a consultant in the document and records management domain for both public and private organizations.

Ruben Verborgh is a researcher in semantic hypermedia at Ghent University – iMinds, Belgium, where he obtained his PhD in computer science engineering in 2014. He explores the connection between semantic web technologies and the web's architectural properties, with the ultimate goal of building more intelligent clients. Along the way, he has become fascinated by linked data, REST/hypermedia, web APIs and related technologies. Ruben is the author of a book on the interactive data transformation tool OpenRefine and several publications on web-related topics in international journals.

A word of thanks

A desire to go beyond academic papers and conference presentations drove us in the summer of 2012 to create the Free Your Metadata project. Initially launching it as a gimmick, we were quickly surprised by the uptake of the learning materials that were put online. As failed musicians, we have found the project to be a great alternative way of getting ourselves on tour across the world. It is truly inspiring to see how many people share our passion for metadata, from the World Bank in Washington to information science students in Addis Ababa. So our thanks first of all go out to all the metadata practitioners we have met, and who made us realize the necessity of this handbook. Second, we also want to warmly thank Max De Wilde, our fellow metadata liberator. Without his skills in computational linguistics and his acting talents in our videos, the project and this book would not have been the same.

Seth specifically wants to thank the following people:

Professor Isabelle Boydens has been from the very beginning of my career until now a true mentor. Her critical thinking and quest for perfection in both research and teaching continuously push me forward. I also wish to thank Professor Eva Méndez and Professor Jane Greenberg, who have both been a great source of inspiration. On a personal level, I want to thank Karen Torres, whose Brooklyn home has been a base camp for establishing contacts in the USA over the years. James Lappin, besides being a great friend, provided critical feedback on the book. Lastly, I want to thank my parents for providing me with all the opportunities I could have wished for to develop myself.

Ruben wishes to thank:

The many people I met over the years for several valuable insights on technology and its use in practice. Thanks to fellow researchers and colleagues for lively discussions. My deepest appreciation goes to my wife Anneleen for being so supportive during all those months of writing. Thanks to my parents for encouraging me to follow my dreams and passions. Finally, a tip of the hat to my musician friends Thomas and Steven, secretly hoping that one day we'll do a tour of the world, too.

Foreword

Never before has so much of our global cultural heritage been at our fingertips. Yet as billions have been spent so far on digitization, both public and private, it still feels as though we are in the very earliest stages of what might be possible. Truly usable and intuitive interfaces notwithstanding, there is still much to do in terms of simple search and discovery tools across multiple collections.

Two of the datasets that feature as case studies in this book are from institutions where I've led the teams responsible for their public release. One, the Powerhouse Museum, was a large and comparatively well resourced state institution, with a long history of rigour, excellence and computer-based documentation amongst its cataloguing and registrar departments; and the other, Cooper-Hewitt, a small historic house museum, whose collection still is best documented on cards dating from the mid 20th century. Even when those two institutions held versions of the same object, such as the seminal Aeron office chair, the two corresponding documentary records could not have been more different in detail and perspective. But at the end of the day, the future of both these institutions lies in what they can do with those records and what can be built upon them now and in a hundred years' time. In making these datasets publicly available, downloadable and accessible through simple well designed APIs, both the Powerhouse and Cooper-Hewitt had to embrace an acceptance of the incompleteness of their digitization and cataloguing efforts. I've previously described this as 'institutional *wabi sabi*', referencing the Japanese aesthetic of imperfection as beauty and applying it to organizational strategy and practice.

This book is a much-needed guide to the 'how' of getting more from those collections that form the backbone of libraries, archives and museums, even if the galleries are now being filled with blockbuster 'experiences', and the stacks replaced with internet terminals and comfortable lounges. In fact, as our knowledge institutions increasingly become places of memorable social experiences and interactions, the importance of collections having greater exposure, access and life to those outside the 'building' is ever more critical.

Importantly, this book introduces the means to see the totality of the process of making your collections available, from the arduous processes of cleaning and connecting to publishing it for the world. Along the way there are diversions into controlled vocabularies, crowdsourcing, APIs, data profiling and code. Often inside institutions these tasks, and associated discourses, are still carried out by individuals located on different branches of the organizational tree, and sometimes even in different buildings. This physical and psychological separation has often contributed to an institutional inertia around making collections available, and my hope is that readers will come to a better understanding of the complexity at each stage of the process and begin to collaborate better. Along with a range of useful case studies and code examples, both the budding digital humanist and the long-tenured registrar and cataloguer should be able to quickly start to poke at and experiment with their own datasets.

If my teams had had access to this book when we started making collections data widely available back in 2006, we might have done things a little differently and better.

Sebastian Chan
Director of Digital & Emerging Media
Smithsonian Cooper-Hewitt, National Design Museum, New York

Glossary

AI	Artificial Intelligence
ANSI	American National Standards Institute
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
ASP	Active Server Pages
BT	Broader Term
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CMS	Content Management System
CSS	Cascading Style Sheets
CSV	Comma-Separated Value
DC	Dublin Core
DDC	Dewey Decimal Classification
DERI	Digital Enterprise Research Institute
DH	Digital Humanities
DNS	Domain Name Server
DPLA	Digital Public Library America
DTD	Document Type Definition
GSC	Gold Standard Corpus
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IANA	Internet Assigned Numbers Authority
IETF	Internet Engineering Task Force
IP	Internet Protocol
IRI	Internationalized Resource Identifier
ISBN	International Standard Book Number
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation

LCC	Library of Congress Classification
LCNA	Library of Congress Name Authorities
LCSH	Library of Congress Subject Headings
LD	Linked Data
LIS	Library and Information Science
LOC	Library of Congress
LOD	Linked Open Data
MACS	Multilingual Access to Subjects
MADS	Metadata Authority Description Schema
MARC	Machine Readable Cataloging
MQL	Metaweb Query Language
NER	Named-entity Recognition
NISO	National Information Standards Organization
NLP	Natural Language Processing
NT	Narrower Term
OCLC	Online Computer Library Center
OCR	Optical Character Recognition
OWL	Web Ontology Language
PNG	Portable Network Graphics
PONT	Powerhouse Museum Object Name Thesaurus
QR	Quick Response Code
RAM	Random Access Memory
RAMEAU	Répertoire d'autorité-matière encyclopédique et alphabétique unifié
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RDMS	Relational Database Management Software
REST	Representational State Transfer
RT	Related Term
SDBM	Schoenberg Database of Manuscripts
SGML	Standard Generalized Markup Language
SKOS	Simple Knowledge Organization System
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SRU	Search & Retrieve via URL
SRW	Search & Retrieve Web Service
STITCH	SemanTic Interoperability To access Cultural Heritage
SWD	Schlagwortnormdatei
TEI	Text Encoding Initiative
TMS	The Museum System
TSV	Tab-Separated Value

UC	Universal Classification
UDC	Universal Decimal Classification
UI	User Interface
ULAN	Union List of Artist Names
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UTF	UCS Transformation Format
VIAF	Virtual International Authority Files
VRA	Visual Resource Association
XML	eXtensible Markup Language

1

Introduction

1 Metadata at the crossroads

While working with metadata practitioners and students over recent years, we often sensed a frustration. Linked data holds the promise to create meaningful links between objects of disparate collections, but the actual implementation tends to be quite complex. According to academics and consultants, RDF triple stores and SPARQL endpoints should bring us a brave new world in which everything is connected in a meaningful manner. Unfortunately, the reality proves to be quite complex and even outright messy. The search for the Holy Grail of data integration can turn into a nightmare, in a world where anyone can state anything about anything. Linked data: the kingdom of structured data to come or an irritating buzzword which we all will have forgotten in a few years?

The ambition of this handbook is to bring a sense of pragmatism to the debate. We will point out the low-hanging fruit currently available, but also identify potential issues and areas where it is uncertain that investments in linked data will deliver benefits. Besides avoiding an uncritical promotion of a hyped technology, the originality of this handbook lies in the positioning of linked data within the broader context of how metadata practices have evolved over the last decades. As many of you may have noticed, the term 'metadata architect' has started to appear on business cards and job descriptions. It is interesting to observe trends in the labelling of professional activities, especially in fields such as information science, which tend to be considered as arcane by the general public. This job title reflects the enthusiasm of the late 1990s and 2000s for the development of metadata standards, application profiles and metadata mappings. The use of the term 'architect' illustrates an almost utopian belief in the design of a global and coherent information architecture which can be implemented consistently across an organization.

Most of us know that the term 'metadata architect' rarely matches the reality. 'Digital landfill manager' sounds less glamorous but reflects the job content more adequately.[1] The implementation of successive technologies over decades has scattered the metadata of our libraries, archives and museums across multiple databases, spreadsheets and even unstructured word processing documents. For business continuity reasons, legacy and newly introduced technologies often co-exist in parallel. Even in cases where a superseded technology is completely abandoned, relics of the former tool can often be found in the content which has been migrated to the new application.

Why does this all matter? This book does not defend a merely technological deterministic view but we do want to emphasize the impact tools have on how we access and use cultural heritage. Throughout different application domains, there is a common belief that the use of new technologies is beneficial. Experience from the terrain has demonstrated that this is unfortunately not always the case. For example, throughout the 1970s and 1980s the first databases were introduced and retro-cataloguing efforts were initiated to convert millions of paper-based cards into database records. Popular database software of that era, such as DBase, ran on hardware with limited storage capacities. Due to this limitation, database administrators restricted the number of characters to be used within fields, leading to a situation where the first databases from the 1980s sometimes contained less detailed information than the paper-based documentation.

As a profession and discipline, we have been working hard over the last two decades to streamline documentation practices in libraries, archives and museums. The rise of the web obliged us to pick up the pace of standardization efforts of metadata schemes and controlled vocabularies, which were initiated after the use of databases for cataloguing and indexing throughout the 1970s and 1980s. At the same time, budget cuts and fast-growing collections are currently obliging information providers to explore automated methods to provide access to resources. We are expected to gain more value out of the metadata patrimony we have been building up over decades. The current hype on linked data seems to offer amazing opportunities to valorize what we already have and to facilitate the creation of new metadata. To what extent can we, as a discipline and a profession, take linked data at face value?

Until recently, metadata practitioners lacked accessible methods and tools to experiment with linked data. In this handbook, we will focus on how freely available tools and services can be used to start evaluating the use of linked data on your own. The quickly evolving landscape of standards and technologies certainly continues to present challenges to non-technical domain experts, but we do want to point out the low-hanging fruit which is currently hanging in front of our noses.

Technology is a means and not an end. Opportunities arise from new technologies, yet never before has it been easier to get lost and trapped within them. Linked data principles are often misunderstood and need to be implemented in a well reflected manner. Linked data present tremendous challenges with regard to the quality of our metadata, and so it is fundamental to develop a critical view and differentiate between what is feasible and what is not.

2 Definition and scope of key concepts

Three concepts define and delimit this handbook: linked data, metadata and cultural heritage institutions. In order to set the expectations right, let us briefly see how these terms are understood within the context of the handbook.

The term **linked data** (often given as 'Linked Data') is often used as if it was a specific, well defined technology. For example, you might have come across technology vendors claiming their products explicitly support linked data. The term does not represent one well defined technology or standard. In the context of this book, linked data is understood as a set of best practices for the publication of structured data on the web. Although a lot of effort is being put into the standardization process of all of the underlying techniques, this set of practices is evolving at a continuous pace. Linked data remains very much a moving target but within this handbook we concentrate on core principles which should remain stable over the years to come.

A linked data handbook with a particular focus on metadata seems to be a tautology, in the sense that linked data as such can be considered metadata. RDF triples are short and simple statements which describe a resource. By doing so, they are data about data and can therefore be considered metadata. This brings up the question of where to draw the line between data and metadata. The short answer is: you cannot. It is the context of the use which decides whether to consider data as metadata or not. You should also not forget one of the basic characteristics of metadata: they are ever-extensible. Just as you can always add an extra Lego piece on top of another, you can always add another layer of metadata to describe metadata. For example, a user review of a book on Amazon can be considered as a form of metadata of the book. By giving users the opportunity to evaluate the usefulness of the review, other users add another level of metadata. The research domain on the issue of provenance and trust on the web is by and large based on this principle. This feature of ever-extensibility comes in very handy but can also turn into a nightmare: every extra layer of metadata adds to the complexity of an application.

Within the context of this handbook, our focus resides on metadata from libraries, archives and museums. Throughout the handbook, we will refer to these as cultural heritage institutions. Each one of these three types of organization has its own deeply rooted traditions regarding metadata. Nevertheless, we think it is useful to synthesize within one handbook common principles and best practices for the management of metadata. When describing practices such as metadata modelling, cleaning, reconciliation, enrichment and publication, the focus of the book will be as much as possible on common needs shared across different types of institutions. However, we do acknowledge the fundamentally different views an archivist, for example, holds when thinking about metadata modelling, compared with those of a librarian or a museum curator. Due to the importance of the notion of the 'fonds', in which the place of an individual document within a larger collection is of central importance, archival collections find a natural fit with the hierarchical tree structure of XML. On the other hand, libraries have worked for decades with the MARC format, which is an electronic file format created in the 1960s to represent flat files containing bibliographic data. These differences have a big impact on metadata practices but throughout the different chapters we have tried as much as possible to develop views and recommendations which are relevant for all three types of institutions.

3 Position and originality of the handbook

Over recent years a wealth of information has been made available on the topic of linked data. In this section we wish to point out the most comprehensive learning resources available but also to emphasize the originality of this handbook when compared to the existing literature.

The computer science community has delivered over the last years specific handbooks on linked data (Heath and Bizer, 2011; Wood, Zaidman and Ruth, 2013). The by now classic semantic web handbook by Allemang and Hendler (2011) is highly relevant for people eager to learn about linked data. Although outdated in some aspects, the practical handbook by Segaran remains a useful resource (Segaran, Evans and Taylor, 2009).

There is to our knowledge no previously published handbook which is aimed specifically at the library and information science (LIS) or the digital humanities (DH) communities. Greenberg and Mendéz published a comprehensive set of chapters on different aspects of the semantic web relevant to the library and information science domain, but the publication remains mainly research-oriented (Greenberg and Mendéz, 2012). People from the LIS community might be surprised by the lack of attention to metadata standardization efforts. Although the topic is relevant in the context of linked data, metadata standards have been abundantly addressed over the last years in other handbooks. Where necessary, we refer to the relevant literature on metadata standards throughout this handbook.

The second chapter of this book is dedicated to metadata quality, which has until now remained under-represented in the linked data literature. The topic of data quality has already attracted attention in other application domains and some handbooks have been devoted to the topic. The most useful book, beyond any doubt, is *Data Quality: the accuracy dimension* by Olson (2003) and is often referred to in this handbook. O'Reilly published an interesting collection of concrete case studies and best practices with the aptly titled *Bad Data Handbook*

(MacCallum, 2012). One of the authors of our book has also published a handbook specifically on the use of OpenRefine, offering readers the opportunity to go into more specific details of all of the functionalities of OpenRefine (Verborgh and De Wilde, 2013).

Compared to other publication formats, handbooks aim to offer a comprehensive introduction to a topic. Within this genre, this handbook specifically has the ambition to:

- lower the technical barrier towards understanding linked data
- propose a critical view of linked data, by not making an abstraction of the challenges and disadvantages involved.

First of all, we wish in this handbook to address the specific needs of people who do not have a technical background in computer science. More particularly, the handbook was written for readers with a background in library and information science and digital humanities. Both communities have shown a strong interest in linked data and hope to leverage through its principles the creation and use of metadata. The two communities have their own tradition and methods with regard to metadata, and it is interesting to bring them together in this book.

Secondly, linked data literature tends to be written by technology evangelists who sometimes hold an almost religious belief in the value of linked data. Unfortunately, technology all too often becomes an end in itself. Based on some of the linked data literature, one might start to think that we will be abandoning our relational databases for triple stores. The reality is that we will continue to use relational databases over the next years (and probably decades), as they excel at managing structured data. Within this handbook, we try to make it very clear what exactly the advantages of linked data are for the publication of your metadata. As with many things in life, advantages often come at a cost. At the end of the day, it is the context of your specific project, with its own needs and resources, that will decide what technology to use. If you can deliver good results with a tool which has existed for over 30 years, then there is absolutely no reason to go along with the most recent technology hype.

4 Structure and learning objectives

In order to achieve the ambitions mentioned above, a lot of effort was put into the structure of the handbook and the combination of theory with practice. This handbook tightly couples the conceptual introduction of technologies with hands-on exercises and experimentation, giving non-IT experts the opportunity to evaluate the practical use of metadata cleaning and reconciliation, named-entity recognition (NER), sustainable publishing and the overall concept of linked data.

Each chapter leads up to a concrete case study with metadata from institutions around the world (USA, Australia and Europe). The accompanying website, http://book.freeyourmetadata.org/, allows you to download the metadata used in the case studies and to repeat the exercises at your own pace.

The chapters and case studies stand on their own and can be read individually. LIS and DH professors and independent trainers can use one of the five core chapters (2–6) to build up a specific class on, for example, metadata cleaning or the use of NER within a more generic metadata or DH-oriented class. One of the biggest incentives to write this handbook was to provide thorough documentation for our own students and workshop participants. We have tested and refined the examples and exercises over the course of three years, with the help of hundreds of students and archivists, librarians and curatorial staff in Europe, the USA and Australia.

Throughout the handbook we have tried to keep as much consistency as possible with regard to the technical skills readers acquire throughout the different chapters. Three out of the five case studies involve the use of OpenRefine. This free and easy-to-use application could be considered as Excel-on-steroids. Visually the interface resembles the popular spreadsheet software, but it offers a host of possibilities for automatically deriving more value out of your existing metadata. Extra functionalities, known as extensions, are constantly being developed for this software. Specifically for the readers of this handbook, we developed an extension which allows the user to apply NER services in a very handy way (see Chapter 5). This extension has been warmly welcomed by the OpenRefine community and is quickly becoming one of the key functionalities for the use of OpenRefine for linked data applications.

Even if the chapters stand on their own, there is a clear logic behind the order in which the chapters are presented. As such, the entire book can also be used as a global handbook on the use of linked data within the humanities. The following list gives a clear outline of the content, learning outcomes and reader profile of each chapter:

- **Chapter 2: Modelling**
  - — Overall goal: understand the rationale of linked data through an overview of the major data modelling paradigms.
  - — Audience: people in need of a better understanding of the differences and similarities between tabular data, relational databases, XML and RDF.
  - — Conceptual insights: impact of data modelling for metadata.
  - — Practical skills: construction of queries in graph databases. Readers are made familiar with SPARQL through DBpedia. In order to make use of Freebase, the proprietary Metaweb Query Language is also illustrated with some examples.

- **Chapter 3: Cleaning**
  - Overall goal: understanding that most metadata need to be cleaned.
  - Audience: collection holders who want to understand how to weed out common metadata quality issues and get a better global understanding of metadata quality.
  - Conceptual insights: quality is a fundamentally relative characteristic; 'total quality' therefore does not exist. Instead, focus on how metadata evolve through time.
  - Practical skills: metadata profiling and cleaning operations with the help of the general features of OpenRefine are illustrated with metadata from the Schoenberg Database of Manuscripts.

- **Chapter 4: Reconciling**
  - Overall goal: possibilities and limitations of re-using controlled vocabularies.
  - Audience: practitioners and students who want to understand the differences between classification schemes, subject headings and thesauri, and how they can be represented in a web-accessible format (SKOS, Simple Knowledge Organization System).
  - Conceptual insights: advantages and disadvantages of the use of controlled vocabularies on the web.
  - Practical skills: after an introduction to SKOS through the manual encoding of a mini-thesaurus with a text editor, the use of the RDF extension for OpenRefine is demonstrated. Once the basic functionalities of SKOS and the creation of reconciliation sources is understood, the case study focuses on how the LCSH can be used to reconcile a collection of metadata records from the Powerhouse Museum.

- **Chapter 5: Enriching**
  - Overall goal: possibilities and limitations of applying NER to metadata.
  - Audience: collection holders who want to understand what types of results can be expected from NER technologies.
  - Conceptual insights: introduction to theme of 'Big Metadata' and applying 'distant reading' techniques upon cultural heritage metadata. Overview of the ambiguity of URLs.
  - Practical skills: step-by-step introduction to the use of the NER extension within OpenRefine. Three different NER services (Zemanta, Alchemy and DBpedia Spotlight) are tested upon the descriptive fields of metadata the British Library has provided through Europeana.

- **Chapter 6: Publishing**
  - Overall goal: understanding how to publish your collection in a sustainable manner.
  - Audience: practitioners and students interested in understanding the conceptual and practical benefits of the representational state transfer (REST) architectural style.
  - Conceptual insights: distinguish resources from their representations.
  - Practical skills: through a small prototype with metadata from the Smithsonian Cooper-Hewitt National Design Museum, readers can experiment with how a RESTful application can publish metadata in a sustainable manner. Exercises with the APIs of Europeana and Digital Public Library America (DPLA) allow the reader to better situate the concepts presented in the theoretical part of the chapter.

5 Get in touch!

Throughout the years, we have learned a lot by talking and working with metadata enthusiasts across the world. As a result we have been able to write this handbook, which is very much an outcome of the discussions we have had with practitioners from libraries, archives and museums. We sincerely hope this handbook will offer us the opportunity to get in touch with even more people. The website http://freeyourmetadata.org/ will be updated with case studies and announcements of workshops and seminars. Do not hesitate to contact us – really. If you have a particularly dirty metadata set you want us to have a look at, get in touch. The dirtier your metadata are, the more we will love them. Or if you are busy developing a global institutional strategy for linked data, we'll be happy to share our thoughts with you. Our research and writing needs your input, so we will be happy to hear from you!

Note

[1] The term 'digital landfill manager' has been proposed by James Lappin in his article on the evolution of records management (Lappin, 2010).

References

Allemang, D. and Hendler, J. (2011) *Semantic Web for the Working Ontologist*, Morgan Kaufmann.

Greenberg, J. and Mendéz, E. (2012) *Knitting the Semantic Web*, Routledge.

Heath, T. and Bizer, C. (2011) *Linked Data: evolving the web into a global data space*, Morgan & Claypool.

Lappin, J. (2010) What Will be the Next Records Management Orthodoxy?, *Records*

Management Journal, **20** (3), 252–64.

MacCallum, E. (2012) *Bad Data Handbook: mapping the world of data problems*, O'Reilly.

Olson, J. (2003) *Data Quality: the accuracy dimension*, Morgan Kaufmann.

Segaran, T., Evans, C. and Taylor, J. (2009) *Programming the Semantic Web*, O'Reilly.

Verborgh, R. and De Wilde, M. (2013) *Using OpenRefine*, Packt Publishing.

Wood, D., Zaidman, M. and Ruth, L. (2013) *Linked Data: structured data on the web*, Manning.

Index