The Taxobook

Principles and Practices of Building Taxonomies

Part 2 of a 3-Part Series

Synthesis Lectures on Information Concepts, Retrieval, and Services

Editor
Gary Marchionini, *University of North Carolina, Chapel Hill*

Synthesis Lectures on Information Concepts, Retrieval, and Services is edited by Gary Marchionini of the University of North Carolina. The series will publish 50- to 100-page publications on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. The scope will largely follow the purview of premier information and computer science conferences, such as ASIST, ACM SIGIR, ACM/IEEE JCDL, and ACM CIKM. Potential topics include, but not are limited to: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

The Taxobook: Principles and Practices of Building Taxonomies: Part 2
Marjorie M.K. Hlava
October 2014

The Taxobook: History, Theories, and Concepts of Knowledge Organization: Part 1
Marjorie M.K. Hlava
October 2014

Children's Internet Search: Using Roles to Understand Children's Search Behavior
Elizabeth Foss and Allison Druin
September 2014

Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security
Edward A. Fox, Ricardo da Silva Torres
March 2014

Digital Libraries Applications: CBIR, Education, Social Networks, eScience/Simulation, and GIS
Edward A. Fox, Jonathan P. Leidig
March 2014

Information and Human Values
Kenneth R. Fleischmann
November 2013

Multiculturalism and Information and Communication Technology
Pnina Fichman and Madelyn R. Sanfilippo
October 2013

The Future of Personal Information Management, Part II: Transforming Technologies to Manage Our Information
William Jones
September 2013

Information Retrieval Models: Foundations and Relationships
Thomas Roelleke
July 2013

Key Issues Regarding Digital Libraries: Evaluation and Integration
Rao Shen, Marcos Andre Goncalves, Edward A. Fox
February 2013

Visual Information Retrieval using Java and LIRE
Mathias Lux, Oge Marques
January 2013

On the Efficient Determination of Most Near Neighbors: Horseshoes, Hand Grenades, Web Search and Other Situations When Close is Close Enough
Mark S. Manasse
November 2012

The Answer Machine
Susan E. Feldman
September 2012

Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach
Edward A. Fox, Marcos André Gonçalves, Rao Shen
July 2012

Online Multiplayer Games
William Sims Bainbridge
2009

Information Architecture: The Design and Integration of Information Spaces
Wei Ding, Xia Lin
2009

Reading and Writing the Electronic Book
Catherine C. Marshall
2009

Hypermedia Genes: An Evolutionary Perspective on Concepts, Models, and Architectures
Nuno M. Guimarães, Luís M. Carrico
2009

Understanding User-Web Interactions via Web Analytics
Bernard J. (Jim) Jansen
2009

XML Retrieval
Mounia Lalmas
2009

Faceted Search
Daniel Tunkelang
2009

Introduction to Webometrics: Quantitative Web Research for the Social Sciences
Michael Thelwall
2009

Exploratory Search: Beyond the Query-Response Paradigm
Ryen W. White, Resa A. Roth
2009

New Concepts in Digital Reference
R. David Lankes
2009

Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation

Michael G. Christel

2009

The Taxobook: Principles and Practices of Building Taxonomies
Part 2 of a 3-Part Series
Marjorie M.K. Hlava

ISBN: 978-3-031-01160-3 print
ISBN: 978-3-031-02288-3 ebook

DOI 10.1007/978-3-031-02288-3

A Publication in the Springer series
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #36

Series Editor: Gary Marchionini, University of North Carolina, Chapel Hill

Series ISSN 1947-945X Print 1947-9468 Electronic

The Taxobook

Principles and Practices of Building Taxonomies

Part 2 of a 3-Part Series

Marjorie M.K. Hlava
Access Innovations, Inc., Albuquerque, New Mexico

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #36

ABSTRACT

This book outlines the basic principles of creation and maintenance of taxonomies and thesauri. It also provides step by step instructions for building a taxonomy or thesaurus and discusses the various ways to get started on a taxonomy construction project.

Often, the first step is to get management and budgetary approval, so I start this book with a discussion of reasons to embark on the taxonomy journey. From there I move on to a discussion of metadata and how taxonomies and metadata are related, and then consider how, where, and why taxonomies are used.

Information architecture has its cornerstone in taxonomies and metadata. While a good discussion of information architecture is beyond the scope of this work, I do provide a brief discussion of the interrelationships among taxonomies, metadata, and information architecture.

Moving on to the central focus of this book, I introduce the basics of taxonomies, including a definition of vocabulary control and why it is so important, how indexing and tagging relate to taxonomies, a few of the types of tagging, and a definition and discussion of post- and pre-coordinate indexing. After that I present the concept of a hierarchical structure for vocabularies and discuss the differences among various kinds of controlled vocabularies, such as taxonomies, thesauri, authority files, and ontologies.

Once you have a green light for your project, what is the next step? Here I present a few options for the first phase of taxonomy construction and then a more detailed discussion of metadata and markup languages. I believe that it is important to understand the markup languages (SGML and XML specifically, and HTML to a lesser extent) in relation to information structure, and how taxonomies and metadata feed into that structure. After that, I present the steps required to build a taxonomy, from defining the focus, collecting and organizing terms, analyzing your vocabulary for even coverage over subject areas, filling in gaps, creating relationships between terms, and applying those terms to your content. Here I offer a cautionary note: don't believe that your taxonomy is "done!" Regular, scheduled maintenance is an important—critical, really—component of taxonomy construction projects.

After you've worked through the steps in this book, you will be ready to move on to integrating your taxonomy into the workflow of your organization. This is covered in Book 3 of this series.

KEYWORDS

taxonomy, thesaurus, controlled vocabulary, search, retrieval, ontology, knowledge organization, classification, theory of knowledge, metadata

Contents

This book is dedicated to all taxonomists, past, present, and future. My team at Access Innovations worked hard and long to bring this book to fruition. It would not have been done without their encouragement, patience, and support.

List of Figures

Preface

Most of us are keenly—personally—aware that over the past several years, information on the Internet has been rapidly expanding, with a flood of information pouring out of computer screens to people everywhere. In 1998, Google reported 3.6 million searches for the year. In 2012, they reported an average of over five billion searches *every day*. That's an increase of over 52 million percent! They claim 67% of the search market, so there remains another 33% of the market of searches to add to that five billion.

We use search often. We use search so often that "Google" has become a verb, at least in practice. "Google it" has become an everyday phrase. Early in my career, searching the Internet (or its precursor, DARPAnet) was the purview of professionals with special training, special access, and special equipment. We were an elite group of gatekeepers, in a way, with access to a corpus of knowledge desirable to researchers but inaccessible except through professional searchers.

In response to our search queries—when we "just Google" something—the search engines like Google, Yahoo, Ask, and others return millions of hits within milliseconds, but how many of those millions of hits does the searcher actually need… or want? How often do you find that the site you seek is at the top of the search results page? How often do you find that the search results don't include what you seek, or that it is buried ten pages down? How often do you look through ten pages of search results to see if your desired site is listed at all? How do we contend with this exploding flood of information and find what we actually need? Search needs help!

A parallel expansion—or explosion—has been occurring in intranets, where individual organizational and enterprise information resides. Organizations are eagerly adopting technologies that can locate and sort out the information that is wanted and needed. In this environment, as Jean Graef of the Montague Institute put it shortly after the turn of the millennium, "Taxonomies have recently emerged from the quiet backwaters of biology, book indexing, and library science into the corporate limelight." Corporate librarians, information technology specialists, and others involved in information storage and retrieval recognize and acknowledge the value of taxonomies. However, these people often lack an understanding of taxonomies and of how they are created, maintained, and implemented.

In response, we have developed this guide to taxonomy creation, development, maintenance, and implementation. We will progress rapidly from theory to practice because both are critical for a comprehensive knowledge. The guide is intended to cover the full spectrum from the original scoping of the work through its use in tagging (indexing with keywords from the taxonomy), web-

site navigation, search, author and affiliation/organization disambiguation, identification of peer reviewers, recommendation systems, data mashups, and a myriad of other applications.

In Book 1 of this three-part series, I introduce the very foundations of classification, starting with the ancient Greek philosophers Plato and Aristotle, as well as Theophrastus and the Roman Pliny the Elder. They were first in a line of distinguished philosophers and other thinkers to ponder the organization of the world around them and attempt to apply a structure to that world. I continue by discussing the works and theories of several other philosophers from medieval and Renaissance times, all the way through to notable modern library science figures, including Saints Aquinas and Augustine, William of Occam, Andrea Cesalpino, Carl Linnaeus, Rene Descartes, John Locke, Immanuel Kant, James Frederick Ferrier, Charles Ammi Cutter, and Melvil Dewey. Part 8 covers the contributions of Shiyali Ramamrita Ranganathan, who is considered by many to be the "father of modern library science." He created the concept of faceted vocabularies, which are widely used—even if they are not well understood—on many e-commerce websites.

I believe that it is important to understand the history of knowledge organization and the differing viewpoints of various philosophers—even if that understanding is only that the differing viewpoints simply exist. Knowing the differing viewpoints will help answer one fundamental question: *why* do we want to build taxonomies?

With that understanding the process will go much faster. Taxonomists must think in a different way from the normal subject matter expert way of thinking. Taxonomy thinking is thinking in interconnected outlines. It is not the strictly linear thinking shown in a single taxonomy or hierarchical view of a taxonomy—that list with its increasing levels of specificity, but rather thinking for many people taking many approaches to a subject. Those who can sit in an ivory tower and pursue a single thread of thought to eventually developing a full outline of knowledge from their point of view will only sever their single point of view. They will have converts to their way of thinking, but they will not support an interconnected search world with each individual looking in from their own unique perspective. But you know how that works from the first volume, so let's really get to the hands-on work.

In Book 2, I suggest reasons for creating a taxonomy and how it can be used to advantage in an organization. I present and describe various forms of controlled vocabularies, including taxonomies, thesauri, and ontologies, and include methods for constructing taxonomies—or other kinds of controlled vocabularies. Standards, especially information standards, are near and dear to my heart, and I have served on several committees and review boards for many of the information standards published by NISO and other standards-forming organizations. Therefore, Chapter 7 of Book 2 provides an abbreviated list of the specific standards that I feel are most important to knowledge and information professionals, brief descriptions of some of the standards-forming organizations, and the process that they go through in creating these standards or guidelines. While standards

might sound like a dry subject best used to cure insomnia, I suggest that they will provide you with an excellent framework for your taxonomy construction project.

Book 3 covers putting your taxonomy into use. It's all well and good to create a beautiful taxonomy that classifies The World as We Know It, that conforms to all of the appropriate standards, and is practically perfect in every way, but what good does it do? In order to get back your investment, you have to integrate your taxonomy into whatever workflow or system your organization employs. In Book 3 we discuss the various ways in which you can apply, implement, and integrate your taxonomy into that workflow, with an emphasis on integrating a taxonomy into search. Lastly, I ponder the future of knowledge management. I don't know exactly where we are going, but I have some good guesses based on where we have been and the trends I see in requests from my clients. Based on my guesses, I provide a few suggestions about areas in which you might start to prepare.

While I can't truly predict the future, I am quite certain that the volume of information coming at us isn't going to go away, lessen in intensity, or slow down. The information explosion is going to continue, and we all need to find ways to make sense of it—to improve retrieval, to refine analysis, to pull out the real value of information so that the people who need it, get it.

I hope that you will find this series practical and useful, and perhaps these volumes will become part of your desktop reference collection. Throughout this series, I attempt to include information that will help you to make a business case for your taxonomy construction project, as well as simple to use, step-by-step instructions for creating a taxonomy and leveraging it in multiple ways throughout your organization.

Acknowledgments

The series started as a series of talks and lectures given to various groups as full-day workshops on how to build and implement thesauri, controlled vocabularies, and databases. The audiences helped hone the messages and poked holes in my assertions when appropriate. This was combined with over 600 engagements over the years with fascinating clients who each needed a similar endpoint but with a unique twist because of their content and their individual visions. These combined with the need to educate staff members in how the work is done and the need to formulate best practices, as well as broad support on the standards bodies, to create an unusual degree of perspective on the knowledge organization and distribution process.

This work would not be possible without the tireless efforts and uncompromising support of many, many people. My business partner and friend for most of my professional life, Jay Ven Eman, has been unstinting in his support and encouragement, although he does occasionally roll his eyes at some of my ideas. The team at Access Innovations, all of whom reviewed the drafts and, in particular, Heather Kotula, Barbara Gilles, Tim Soholt, and David (Win) Hansen,who massaged the drafts, untangled my prose, improved the images and examples, and offered very pertinent suggestions to create the final product. Our customers for providing the content and allowing us to work with it have provided an unparalleled laboratory of material for organization to meet their individual needs. To my own family for their cheerful understanding and putting up with the demands of career and writing, my husband Paul Hlava, my daughters Heather and Holly and their families. And to my mom, Mary Kimmel, who showed me that you can have a career and a happy family too.

Many people encouraged me to write down what I was teaching, and I am grateful for their continued insistence. Tim Lamkins for his early review and insightful comments, clients whose works we reference in case studies and examples, and my industry mentors including Roger Summit, Eugene Garfield, Buzzy Basch, Tom Hogan, and Kate Noerr.

To all of these and more I thank you; I could not have done this without you!

Marjorie M.K. Hlava