**SPECIAL SECTION**

**TAXONOMIES IN PRACTICE**

“As technologies, platforms and uses have developed, information specialists have discovered what works and what doesn’t and have developed effective approaches for creating taxonomies that do.”

Page 11

11] **Introduction**  
by Marjorie M.K. Hlava

13] **Evaluating Classification Schema and Classification Decisions**  
by Denise Bedford

22] **Case Study: Developing the PLOS Thesaurus**  
by Jonas Dupuich and Gabe Carr

26] **Building User Experiences: Synchronizing User Experience Design and the Supporting Metadata and Taxonomy Infrastructure**  
by Carol A. Hert, Gary Carlson and Bram Wessel

30] **Indexing Electronic Medical Records Using a Taxonomy**  
by John Kuranz and Barbara Gilles

34] **Building a Taxonomy for Auto-classification**  
by Wendi Pohs

39] **Building Controlled Vocabularies for Metadata Harmonization**  
by Marcie Zaharee

---

**DEPARTMENTS**

[2] Editor's Desktop

[3] President's Page

[3] 2012 President Diane Sonnenwald

[6] 2013 President Andrew Dillon

[8] Inside ASIS&T

**COLUMNS**

[43] IA: On What Assumptions Do You Base Your Thinking? Contextual Challenges and Strategic Possibilities  
by Thom Haller

[46] RDAP: Digital Curation Postdoctoral Fellowship – The Hybrid Role of Data Curator  
by Inna Kouper
Taxonomies are the focus of this issue and, in particular, the practical problems associated with creating, integrating and maintaining these crucial components of many information retrieval systems. Marge Hlava of Access Innovations, Inc., is our guest editor and has recruited timely articles on a variety of topics ranging from taxonomy creation and evaluation to integrating taxonomy structure into user experience design and utilizing taxonomies in automatic indexing. Our thanks to Margie for her efforts.

The issue includes our regular IA column and the RDAP Review column that was initiated as a regular feature in the last issue. Thom Haller, our associate editor for IA, suggests ways to elevate IA so that it can become more a part of the strategic thinking of management, reducing the risk that higher management will chase after the latest fashionable technology solution regardless of its suitability to their application. It’s an important, ubiquitous problem, and Thom offers sound advice.

Turning to RDAP, the newly established Digital Curation Postdoctoral Fellowship is sponsored by the Council on Library and Information Resources’ Digital Library Federation (CLIR/DLF), the Alfred P. Sloan Foundation and six universities. One of its first recipients, Inna Kouper, is the author of this month’s column, which focuses on the hybrid role of the data curator.

Finally, we have two President’s Pages in this issue: one from outgoing president Diane Sonnenwald and the other from our president for 2013, Andrew Dillon. Together they provide an excellent overview of the strategic direction of the Society. While Andrew’s column is based on his inaugural address at the Annual Meeting, our primary Annual Meeting coverage will be in the next issue. I was, unfortunately, among the missing this year, but I’m told this 75th Anniversary event was a unique experience for the attendees and successful in the face of the challenges presented by superstorm Sandy.
Dear Colleagues,

Writing this President’s Page in early October as my term winds down, I want to use the opportunity to revisit some of the activities we undertook during the past year and to highlight discussions by the ASIS&T Board of Directors. Subsequent activities by the Board and members will be covered in the President’s Page of incoming president Andrew Dillon.

The Board met in August to review our finances and activities and to consider ways forward for our organization. ASIS&T’s finances continue to be strong under the leadership of Vicki Gregory, ASIS&T treasurer, and Dick Hill, our executive director. Task force and committee volunteers have been working on a variety of projects and deserve our thanks for donating their time and energy to make our organization successful. Their recent accomplishments are highlighted below. In looking to the future, the Board discussed ASIS&T’s name, membership categories and benefits, the Annual Meeting, partnering with other conferences and possible future collaborations with other organizations. Summaries of those discussions are on the following pages.

EDITOR’S SUMMARY

Closing her year as ASIS&T president, Diane Sonnenwald looks back with thanks to the many who volunteered time and effort as committee and task force chairs and members. Their accomplishments span promoting the discipline and advocating for information professionals, reaching beyond national boundaries through social media and meetings, and establishing an investment program to enhance finances. Committees also managed the awards process, expanded continuing education offerings and coordinated with ISO and NISO on standards. A name change will be voted on by the membership to recognize the association’s international scope. Membership categories are under review, and organizational collaborations and conference partnerships are being examined, all to promote the discipline and the association.

KEYWORDS
information associations
information professionals
continuing education
international aspects
collaboration
meetings
honors
marketing

Diane Sonnenwald was 2012 ASIS&T president. She is head of school and professor at the School of Information and Library Studies at University College Dublin and an adjunct professor in computer science at the University of North Carolina. Prior to joining academia she worked at Bellcore. The National Library of Medicine, National Institutes of Health, National Science Foundation and European Science Foundation, as well as private corporations and foundations, have funded her research. She can be reached at diane.sonnenwald<at>gmail.com.
Task Force and Committee Activities

First of all, I’d like to thank all committee and task force chairs and members for their efforts during the past year.

The Information Professionals Task Force, led by Sandy Hirsh, Prudence Dalrymple and Marcia Bates, is creating plans to strengthen our efforts to advocate for our discipline. Their plans include creating a first draft of a website to help market and advocate for information professionals.

The 75th Anniversary Task Force, led by Toni Carbo and Bob Williams, planned – and then implemented – an interesting pre-conference workshop and website celebrating our 75th anniversary. For those whose Annual Meeting plans changed with Superstorm Sandy’s arrival, you can read the interesting historical information on the website at www.asis.org/asist2012/historyofASIST.html.

The Publications Committee, chaired by Bob Allen and Neil Smalheiser, has worked this year with Melissa Weaver at ASIS&T headquarters to increase ASIS&T’s presence in social media, while the International Relations Committee, chaired by Mei-Mei Wu, is investigating the feasibility of holding an Annual Meeting outside North America. Thanks to all members who completed their recent survey.

The Budget and Finance Committee, led by treasurer Vicki Gregory, researched and established an investment program that was approved earlier by the Board. This program will enable ASIS&T to earn income from its cash reserves.

The Standards Committee, chaired by Mark Needleman, has continued to expertly review ISO and NISO standards and make recommendations regarding these standards on ASIS&T’s behalf.

The Awards & Honors Committee, chaired by Katharine McCain and Ann Prentice, expertly arranged the many awards review processes, while the Nominations Committee, chaired by Linda Smith, selected an excellent slate of candidates for the membership to consider.

The Leadership Committee, led by Iris Xie, worked to create the leadership workshop at the Annual Meeting and review the applications for the James M. Cretos Leadership award.

The Information Science Education Committee, led by Kathleen Burnett and June Abbas, and the Online Education Task Force, chaired by Diane Rasmussen Neal, supported interesting Webinar offerings and explored future online education opportunities for ASIS&T.

Donald Case, leading the Constitution and Bylaws Committee, and Steve Hardin, serving as Parliamentarian, kept us all well informed about our rules and regulations.

Thanks again to all who worked on the task force groups and committees. All members appreciate your willingness to serve and your contributions. Your efforts help other members and our discipline in countless ways.

Looking Toward the Future... Board Discussions in August

Vote Regarding Our Name. The Board voted to have a vote regarding our name, specifically, whether ASIS&T should become the ASSociation for Information Science & Technology instead of the American Society for Information Science & Technology. This issue was discussed in an earlier Bulletin (see www.asis.org/Bulletin/Apr-12/AprMay12_PresidentsPage.html) and in a Quicktopic discussion list (see www.quicktopic.com/47/H/bvJVhSC8HTs). All active members have received a ballot.

Exploring New Membership Categories and Benefits. The Board also discussed membership categories, and we will be exploring new forms of institutional memberships, initially focusing on what membership benefits might encourage information science educational programs/schools to join ASIS&T.

Improving Annual Meeting Processes. Previously, co-chairs of our Annual Meeting had to pay a registration fee in addition to working hard for over a year to organize the Annual Meeting. The Board voted to provide free registration to the co-chairs as a small token of our appreciation for their efforts.

The Board also established a subcommittee, consisting of...
Diane Neal, Katriina Byström and Andrew Dillon, to create an Annual Meeting guidelines document to assist future co-chairs and headquarter staff. No such documentation exists currently.

The Board further established a second subcommittee, consisting of Cassidy Sugimoto, Shelly Warwick and Dick Hill, to investigate how workshops at the Annual Meeting might better meet the needs of chapters and SIGs.

Partnering with Other Conferences. The Board has voted to partner with the Library 2.012 (www.library20.com/page/2-012-conference) and the International Symposium on Information Science (ISI2013). Last year the Library 2.011 Conference had 6800 people, who signed up from more than 150 countries. ISI2013 (www.isi2013.de) is the main conference for information science in German-language countries, attracting 200-300 participants. The Board voted to be a no-cost partner for each of these conferences. This type of partnering helps increase ASIS&T’s visibility in communities that overlap with our own with no financial obligation.

Possible Collaborations with ALISE and iSchools. The current and incoming presidents of ALISE, the iSchool Caucus and ASIS&T recently held a conference call to discuss possible collaborations among our organizations in the future. During the call we brainstormed regarding possible areas of collaboration, including advocacy for the information professions. We hope these discussions will continue.

In Conclusion, Thanks to All

Cassidy Sugimoto (IU) has again worked hard this year on behalf of ASIS&T, serving as Chapter Assembly Chair and member of the Board. Her term ends with the Annual Meeting in October. Other Board members whose terms end this year are Elaine Toms and Prudence Dalrymple. Thank you very much for your efforts.

I also wish to thank the Annual Meeting co-chairs, Crystal Fulton, ShanJu Chang and Julie Hersberger, as well as the many track co-chairs. Few who have not co-chaired a conference of this size realize just how much work is involved in planning a meeting that is not only enjoyable by all, but also inspiring to its audience through the various workshops, presentations, meetings and networking events.

My thanks also to all Board members, including Linda Smith, immediate past president, and Andrew Dillon, incoming president, and to ASIS&T staff. It’s been a pleasure working with everyone this year, and I look forward to serving as immediate past president in the upcoming year.
The occasion of an inaugural presidential address is potentially daunting, representing as it does the moment when the suffix “elect” is removed, and you finally recognize that you are the person tasked with leading the association. A quick scan of previous holders of this office can add to the sense of unease as you see the names Watson Davis, Harold Borko, Eugene Garfield, Clifford Lynch and so many others; it gives one pause—and time to reflect that you also might just be getting a little older than you care to admit.

Any professional society moves slowly but it is the responsibility of the president to make sure it moves surely. With a presidential term being only one year, it is important that collective efforts over years build on what goes before, and in this regard, I intend to build on the efforts of Diane, my immediate predecessor, and those before her who initiated a push toward greater internationalization in order to build strength through a broader, global membership. It is clear that the energy and enthusiasm of our international members are key building blocks of a secure future for the association and, moreover, the concerns of an association claiming to be “the information association for the information age” are truly global. Our name and our actions should recognize this.

I certainly want us to bring new energy to our student chapters by offering resources, wherever possible, to enable chapters to organize and to become involved in the international...
The information field is often split into many competing groups and organizations. I want ASIS&T to build bridges across the divisions so as to live up to our claim of being the information society.

In the coming year we will examine all forms of involvement in ASIS&T so as to engage as many members as possible in the major events of the society: the Annual Meeting, summits, webinars and so forth. To this end, the board is engaging this year in a review of the Annual Meeting structure and program content to ensure the most relevance to members. We need to move beyond practices of habit and avoid undue shift based on the makeup of a particular year’s program committee. The goal is to develop a clearer sense of how the conference can best serve the Society over time. In the same vein, we intend to explore our entire range of online communications, from the website to our continuing education seminars, with a view to developing a consistent and content-rich series of offerings that add real value to members.

But hand in hand with such continued efforts, I believe we need to think clearly about the role of our professional organization and reimagine what ASIS&T can offer to members in the 21st century. I do not just mean analysis of the benefits, the products of membership such as the pricing of the Annual Meeting or access to a digital library. I am thinking more about what membership of this professional association really gives you. I believe that the sense of community and identity that is fostered by membership might seem intangible, but, on reflection, it is key to a positive membership experience. Further, if we cannot easily demonstrate the value of ASIS&T to its members, then why exist and why ask others to spend their money to join?

I am sure many of you, like me, view ASIS&T as being unique in its coverage of the full information field, beyond agencies or job titles, and in so doing we put user values at the core of our mission. Are we clear ourselves on how and why we take this approach? This year I want to help members find better ways to tell our story and to articulate the aims of ASIS&T so that we can help shape better understanding of information education in our universities and of the value in information research to funding agencies and help our graduates and fellow professionals best explain their contributions to employers.

The information field is often split into many competing groups and organizations. I want ASIS&T to build bridges across the divisions so as to live up to our claim of being the information society. We must engage with relevant other groups including the iSchools and ALISE. It is worth recognizing that our membership is dropping while the iSchools movement is growing: so what can we learn here? Is this not a time when the growth in information education should be leveraged for mutual benefit? When Harry Bruce takes over from me next year, he will become the 5th ASIS&T president in a row to come from the ranks of the iSchools, so let us exploit the potential for constructive partnership.

There is work to be done, and it will take more than one year to make real progress, but let us accept the challenge of creating the best information association, member by member, chapter by chapter, SIG by SIG, so as to demonstrate the very leadership mantle we claim.
While hurricane Sandy battered cities up and down the Atlantic coast in late October, ASIS&T loyalists hunkered down in Baltimore to celebrate their Society’s 75th anniversary at the 2012 Annual Meeting.

“We had some no-shows because of the weather,” said Dick Hill, executive director, “but for the most part people came to town and we continued with our Annual Meeting as planned.”

The final count indicated that perhaps 80 people who had pre-registered didn’t make it to the meeting, but for the 85% who did arrive, as well as the nearly four dozen who came and registered onsite, the 75th anniversary celebration of the American Society for Information Science and Technology was an exciting substance- and fun-filled success.

The Annual Meeting activities began with two days of pre-conference seminars and workshops. Among these was a special ASIS&T 75th Anniversary Event entitled History of ASIS&T and Information Science and Technology Worldwide, featuring a keynote address by W. Boyd Rayward, professor emeritus in the Graduate School of Library and Information Science at the University of Illinois. Organized by the 75th Anniversary Task Force, co-chaired by Toni Carbo and Robert Williams, the session featured more than a dozen papers from authors around the world covering such topics as the evolution of the field of information science, historical contexts of technology innovations and impacts, and development of foundational ideas and theories in information science. Nearly 100 attendees were in the audience for this significant session.

Making Lemonade

Though for the most part, the ASIS&T 75th Anniversary Annual Meeting weathered the storm that brutalized neighboring communities, a few sessions were cancelled when participants were unable to make it into Baltimore. But recognizing an opportunity to convert misfortune into lemonade, ASIS&T will turn at least one of the cancelled sessions into a webinar for ASIS&T members and the information community-at-large. Details will be forthcoming, but look for news at the ASIS&T website and in future issues of the Bulletin.

Annual Meeting Coverage

In keeping with recent tradition, our photographic and substantive coverage of the 2012 ASIS&T Annual Meeting will be included in the February/March 2013 issue of the Bulletin. But here, you’ll find a list of the winners of the 2012 ASIS&T Annual Awards for which more details will be provided in the next issue.

ASIS&T Celebrates Its 75th; Ignores Unexpected Guest Hurricane Sandy

Award of Merit: Michael Buckland, University of California at Berkeley
Watson Davis Award: K.T. Vaughan, University of North Carolina at Chapel Hill
Research in Information Science Award: Kaleervo Järvelin, University of Tampere, Finland
Best Information Science Book Award: Modern Information Retrieval by Ricardo Baeza-Yates and Berthier Ribeiro-Neto
Thomson Reuters Outstanding Information Science Teacher Award: Dietmar Wolfram, University of Wisconsin-Milwaukee
James Cretsos Leadership Award: Naresh Agarwal, Simmons College
Pratt Severn Best Student Research Paper Award: April Lynne Earle, St. John’s University, “Design of an Application Profile for the St. John’s University Oral History Collection”
Thomson Reuters Doctoral Dissertation Scholarship: Lori McCay-Peet, Dalhousie University, for “At the Intersection: Investigating the Qualities of the Serendipitous Digital Environment and the Serendipity-Prone Person”
ProQuest Doctoral Dissertation Award: Jaime Snyder, Syracuse University, “Image-Enabled Discourse: Investigating the Creation of Visual Information as Communicative Practice”

Continued on next page
Former ASIS&T president Nancy Roderer, director of the Welch Medical Library and the Division of Health Sciences Informatics (DHSI), will retire, effective January 15, 2013. Also a professor in the Schools of Medicine and Public Health, Nancy has been at Johns Hopkins since 2000. Nancy was instrumental in creating an educational program at the School of Medicine, which includes a Ph.D. in health sciences informatics, two master’s degrees and three certificate programs. Nancy will join her husband David in retirement and will continue to participate selectively in teaching and research projects.

Miles Efron, assistant professor in the Graduate School of Library and Information Science at University of Illinois at Urbana-Champaign, is principal investigator on a half-million dollar grant from the National Science Foundation for a three-year project to investigate the role time plays in information retrieval, specifically in the use and functionality of search engines. In addition, Miles was named featured alumnus by the Alumni Association Board of the School of Library and Information Science, University of North Carolina at Chapel Hill.

The National Science Foundation (NSF) has awarded Rutgers researcher Paul Kantor nearly $1.0 million as part of an initiative to extract useful information from so-called “big data”—massive collections of data from sources such as scientific documents, orbiting instruments, digital images, social media streams and business transactions. The Rutgers grant is part of a $3 million research effort in collaboration with Cornell and Princeton to improve the accuracy and relevance of complex scientific literature searches.

Kantor is a professor in the Department of Library and Information Sciences in the School of Communication and Information, where he heads the LAIR Laboratory. The researchers from the three institutions have been working together for over a year and are putting the finishing touches on the experimental system that will be used in the research.

SIG/Classification Research (SIG/CR) in collaboration with Information Today, University of Washington Libraries Digital Initiatives and the graduate assistants at the University of Washington Information School announces the launch of ACRO: Advances in Classification Research Online, a new web-accessible collection of the proceedings from the annual ASIS&T SIG/Classification Research Workshops. ACRO is indexed in Google Scholar and in Elsevier indexing products.

SIG/Social Informatics (SIG/SI) has announced plans to publish Social Informatics: Past, Present and Future, featuring papers presented at the SIG’s 8th Annual SI Research Symposium at the
ASIS&T 75th anniversary meeting in October. The symposium offered 10 presentations and several posters. The published volume will be edited by symposium organizers, Pnina Fichman and Howard Rosenbaum, both of the School of Library and Information Science at Indiana University.

SIG/Information Policy (SIG/IP) is coming back to life after a somewhat dormant period. Reinvigorated with new leadership, SIG/IP will continue its mission dedicated to awareness of legislation and regulation affecting the information field and/or members of the information professions. It provides venues which encourage research on and discussion of current and emergent information policies as they affect society. It will promote awareness of state/provincial, national, and international policies of public and private entities. With plans on tap for next year’s activities, the group is seeking additional members as well as volunteers for its social media/communications officer and for a webmaster.

SIG/IP is interested in both social repercussions of information policy and the logistics of the legal aspects. Information policy is a part of every special interest group within ASIS&T. Records and knowledge managers have rules, regulations and cumulative knowledge to manage; international information issues teem with information policy problems; and the health care industry is subject to very detailed laws regarding use and privacy of health information. Every aspect of our lives, our world, is affected by information policy. SIG/IP is the place for individuals interested in these and other related topics, as well as for SIGs seeking collaborative projects on information policy issues.

Contact Brandi Loveday, current chair, at sig.ip.chair<at>gmail.com.

SIG/Information Needs, Seeking and Use (SIG/USE) has announced the winners of its annual awards in a variety of categories, as follows:

**Best Paper Award:** June Ahn, Mega Subramaniam, Kenneth R. Fleischmann, Amanda Waugh.

**Best Poster Award:** Ji Yeon Yang and Soo Young Rieh for Dual Roles in Information Mediation at Work: Analysis of Advice-receiving and Advice-Providing Diary Surveys

**Chatman Award:** Joung Hwa Koo, Yong Wan Cho and Melissa Gross for Is Ignorance Really Bliss?: Understanding the Role of Information-Seeking in Coping with Severe Traumatic Stress Among Refugees

**Student Travel Award:** Laura Christopherson

**Interdisciplinary Travel Award:** Roberto Gonzalez-Ibanez

**Outstanding Contribution to Information Behavior Research:** Professor Pertti Vakkari, for his work on task-based information searching and information seeking strategies that have made significant contributions to the field of information behavior research, bridging the interactive information retrieval and information behavior fields.

Greg Walsh and Allison Druin for Youth Identities as Remixed in an Online Community of Storytellers: Attitudes, Strategies and Values

The International Calendar of Information Science Conferences (icisc.neasist.org/) is a nonprofit collaboration between the Special Interest Group / International Information Issues (SIG/III) and the European (ASIST/EC) and New England (NEASIST) Chapters of the American Society for Information Science and Technology, with the additional support of Haworth Press.
Introduction
by Marjorie M. K. Hlava, Special Section Guest Editor

Over the past few decades, taxonomies have grown in importance in the worlds of information technology and knowledge management. As technologies, platforms and uses have developed, information specialists have discovered what works and what doesn’t and have developed effective approaches for creating taxonomies that do. The articles in this issue focus on taxonomy development and implementation.

Perhaps you already have a taxonomy or classification scheme. So how do you evaluate it and decide if it does what you want it to do and what still needs to be done to it? This is the subject of “Evaluating Classification Schema and Classification Decisions,” by Denise Bedford, a well-known taxonomy expert and currently the Goodyear Professor of Knowledge Management at Kent State University. She offers practical methods for evaluating not only the scheme itself, but also how well it supports classification decisions.

Taxonomies, including thesauri, have become important resources in online science publication platforms. For their full potential to be realized, they need to be carefully designed and developed. In “Case Study: Developing the PLOS Thesaurus,” Jonas Dupuich of the Public Library of Science and Gabe Carr of Access Innovations, Inc., describe the rebuilding of the PLOS thesaurus to increase its usability and better reflect the content of its publications in a wide variety of subjects in medicine, biology and other scientific areas.

“Building User Experiences: Synchronizing User Experience Design and the Supporting Metadata and Taxonomy Infrastructure” is by Carol Hert and Gary Carlson of Gary Carlson Consulting and Bram Wessel, an experience strategist at Factor. They observe that there is often a disconnect between...
user experience design work and the development of taxonomy and metadata. The article examines some possibilities for integrating the “dual development processes,” drawing on examples from the authors’ recent projects.

One of the most promising areas for applying digital taxonomy-based platforms is health information. In “Indexing Electronic Medical Records Using a Taxonomy,” John Kuranz, CEO of Access Integrity, and Barbara Gilles of Access Innovations explore some of the challenges involved in achieving indexing accuracy. As the article puts it, “Part of the grand mission of rendering order out of chaos is to bring clarity and precision to the language of our deliberations.” Some automated and semi-automated methodologies for accomplishing this goal are discussed.

In the not-too-distant past of online taxonomies, they were used largely to provide a basis for online navigation of websites. Wendi Pohs of InfoClear Consulting points out that while taxonomies still fulfill this role, the emphasis has been shifting toward their use in auto-classification. This change calls for a change in the taxonomy building process. In “Building a Taxonomy for Auto-classification,” Pohs explains how “we still work to create high-level categories, but we also work to accurately associate the lower-level entities with the higher-level categories we’ve created.”

“Building Controlled Vocabularies for Metadata Harmonization,” by Marcia Zaharee, lead information systems engineer at MITRE Corporation, is a use case for metadata harmonization. The article describes MITRE’s Tactical ISR Integration Metadata Harmonization project for the U.S. Department of Defense. A practical, step-by-step controlled vocabulary or taxonomy development approach is provided, along with the caveat, “Taxonomists will tell you there is no right way to create a CV and taxonomies are always evolving. This is true!”

My thanks to all the authors for sharing their insight and experience.
Evaluating Classification Schema and Classification Decisions
by Denise Bedford

EDITOR’S SUMMARY
Direction on the construction and application of classification schemes such as taxonomies is readily available, but relatively little has been offered on evaluating the schemes themselves and their use to categorize content. A classification scheme can be judged for how well it meets its purpose and complies with standards, and a strong evaluative framework is reflected in S.R. Ranganathan’s principles of classification. The degree of certainty of classification decisions depends on objective understanding of the object to be classified, the scope and details of the class and the coverage and organization of the overall classification scheme. The more complete the information about each class, the more reliable the goodness-of-fit for an object to a class is likely to be, whether chosen by human or machine classifiers. This information comes through definitions, examples, prior use and semantic relationships. The risk of misclassification can be reduced by analyzing the goodness-of-fit of objects to classes and the patterns of missed or erroneous selections.

KEYWORDS
classification schemes, indexing
taxonomy, manual indexing
evaluation, automatic classification

Special Section

This article considers how evaluation pertains to taxonomies. Taxonomies and evaluation are both rich concepts, so it is best to start out with some definitions that help to define our discussion. What do we mean by taxonomy? And what do we mean by evaluation?

Taxonomies
For seasoned information professionals the traditional characterization of a taxonomy is as a hierarchical classification scheme. This characterization has expanded in the last 20 years as the taxonomy community and the information environment have expanded. Today the taxonomy community includes people who design taxonomies, those who build systems that support them and those who use them. Our complex information environment may call for a variety of taxonomic structures, including

- flat taxonomies such as lists of languages or lists of countries;
- hierarchical taxonomies such as topical or subject classifications, business classifications or service classifications;
- faceted taxonomies such as metadata or parametric search structures;
- ring taxonomies such as synonyms or authority control data; and
- network taxonomies such as fully related thesauri or knowledge networks.

Each of these structures has its own set of principles and behaviors. And each requires an evaluation method that aligns with those principles and behaviors. This article focuses on the second type of taxonomy – the traditional classification scheme or hierarchical taxonomy. Classification schemes govern the organization of objects into groups according to explicit properties or values. Classification schemes are in widespread use in everyday life – from grocery stores to websites to personal information spaces.
**Evaluation**

To evaluate something is to determine or fix a value through careful appraisal. There seem to be two important evaluation points related to classification schemes. The first is an evaluation of the classification scheme itself. The second is how well the scheme supports classification decisions. Each requires its own framework and context.

**Evaluating a Classification Scheme**

We can evaluate a classification scheme based on its intended goal and purpose and by how well it aligns with professional standards and principles. Goals and purpose will be institution-specific and are best addressed internally by those who design and work with the scheme. Evaluating a classification scheme according to professional standards and principles, though, is a process that can be generalized. ISO 11179-2 Information Technology – Metadata Registries (MDR). Part 2. Classification (2005) [1] provides advice for constructing data structures and relationships that are used to represent a scheme. ISO 25964 Thesaurus Schemas [2] and ANSI-NISO Z39.19 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies [3] provide some guidance on the distinction between a thesaurus (a network taxonomy structure) and a classification scheme (a hierarchical taxonomy). Important sources for advice on how to construct classes in a classification scheme, though, derive from other sources such as S.R. Ranganathan’s *Prolegomena to Library Classification* [4] and discussions of set theory found in the mathematical sciences literature [5]. Table 1 provides a sample set of principles for constructing classes derived, but reinterpreted, from Ranganathan for an information technology team. The challenge with using these sources for evaluation is that they generally require substantial translation to be understandable by the teams that are building and appraising the scheme. A full set of interpreted principles is available from the author upon request.

It has generally been my experience that Ranganathan’s principles align with but are more exhaustive than the popular guidelines found in the usability engineering literature. The challenge, though, is that they are difficult to interpret by anyone outside of the information science profession. Adapting and interpreting Ranganathan’s principles will provide you with a very strong framework for evaluating the strength of your classification scheme. In fact, the principles convert very nicely into a working checklist for periodic evaluations.

<table>
<thead>
<tr>
<th>Ranganathan’s Principles</th>
<th>Institutional Reinterpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusiveness, Uniqueness, Relevance</td>
<td>No two categories should overlap or should have exactly the same scope and boundaries. Each category should have a set of concepts, derivation rules and content that define its focus and that should be relevant to the goal of the scheme.</td>
</tr>
<tr>
<td>Ascertainability</td>
<td>Each category should be definitively and immediately understandable from its name.</td>
</tr>
<tr>
<td>Consistency</td>
<td>The rules for creating, managing and retiring categories should be consistently adhered to across application systems, across types of information and in different contexts.</td>
</tr>
<tr>
<td>Affinity, Decreasing Extension, Context</td>
<td>Each category in a scheme should be defined in the light of its parent category. While moving down a hierarchy from top to bottom, the intention of the categories should increase, and the extension of the categories should decrease.</td>
</tr>
<tr>
<td>Currency</td>
<td>Names of categories in a classification scheme should reflect the language of the knowledge domain.</td>
</tr>
<tr>
<td>Differentiation</td>
<td>When differentiating a category, it should give rise to at least two subcategories. However, not all levels in the scheme must be differentiated.</td>
</tr>
<tr>
<td>Exhaustiveness</td>
<td>The categories in the scheme should be exhaustive of their common immediate universe; in other words they should provide comprehensive coverage of the knowledge domain of interest.</td>
</tr>
</tbody>
</table>
Evaluating Classification Decisions

Evaluating classification decisions is less straightforward. In order to evaluate a classification decision, we need a good description of the classification process. Within the process description we can pinpoint what to evaluate. Classification is a decision-making process that involves making choices. Typically, choices are made in the context of an existing classification scheme by a human or machine classifier and for a given object (Figure 1).

In theory, this choice seems like a straightforward decision process. The classifier who knows the classification scheme, considers what is known about the object, what is known about the possible classes and makes a decision as to which class in the scheme is the best fit for the object.

The research on classification evaluation is extensive. The citations in this article are illustrative and not comprehensive. The research tends to focus on several contexts for evaluation, including
- comparison of human and machine classification practices [5, 6, 7, 8, 9, 10];
- assessment of the variability of classification decisions among human and machine classifiers [11, 12, 13];
- comparison of machine-generated classification structures and well-established classification schemes and thesauri [14, 15, 16, 17, 18];
- the quality of classification in the context of information retrieval [19, 20, 21]; and
- evaluations of the quality of statistically generated classes [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36].

One fundamental perspective appears to receive less treatment – the simple question of how well the object fits a class in a classification scheme. We suggest there are two simple reasons why this perspective has not received more attention. First, to date most classification is done by people, and we have always assumed that humans make optimal decisions. Second, until recently we have not had the capacity to evaluate decisions in a direct and controlled way. Rather, we have had to evaluate them from an information-retrieval and end-user perspective. What would a direct evaluation of the fit between an object and a chosen class look like?

In an expanding universe of information, classification decisions may be made by people who have neither professional information science training nor subject expertise. Classification decisions may also be made by machine classifiers. Regardless of who makes decisions, the goal is to ensure that those decisions are optimal. An optimal classification decision reflects the best choice that can be made given the information available at the time. An optimal choice may be defined as a good fit between the object and possible classes. An optimal decision also reduces the risk of misclassification. Misclassification may take two forms. The first form occurs when we assign the object to a class for which it is not a good fit. In this case, the object will be presented to the user in error. The second form occurs when we fail to assign the object to a class for which it is a good fit. In this case, the object will be overlooked because it is not in the class. So our evaluation point for classification decisions is determining how well the classified object aligns with the chosen class(es).

Reducing Uncertainty in the Classification Decision

Information economists tell us that optimal decisions result from reducing uncertainty. One way to improve a classification decision, then, is to reduce the uncertainty in the process. Classification is characterized by several kinds of uncertainty. The classifier may have an incomplete understanding of the object. Uncertainty may be high where the classifier has access only to an abstract or summary of the object. The classifier may be uncertain as to what properties or attributes define the class. Uncertainty about the class may result from an incomplete understanding of the scheme or an incomplete specification of the domain – perhaps not all relevant classes have been defined.
in the schema. Perhaps the classifier has imperfect knowledge of all the topics covered in the scheme. The scope and coverage of the classes may not be explicitly available, requiring subjective interpretation by the classifier.

Any of these uncertainties may lead to a suboptimal classification decision. In some cases, making a less than perfect classification decision may be acceptable – perhaps the risk resulting from a classification decision is not so great. If a young reader overlooks a book about the role of snakes in a desert ecosystem for a school project because it was misclassified, the risk is low. In other cases, though, the risk of misclassification may be significant. Where an energetic-materials scientist overlooks an important report of a chemical experiment, national security risks may arise.

These uncertainties are important when we are making a classification decision and when we are evaluating a classification decision. How can we reduce the uncertainty we find in the classification decision? How can we improve the information we need to evaluate classification decisions? The answer is simple. By expanding the information we have about the object, individual classes and the overall makeup and purpose of the classification scheme. Table 2 identifies some of the conditions that might produce low, moderate and high rates of risk in classification decisions.

Notice that the more explicit and objective information we have about each of the factors, the lower the uncertainty. Uncertainty is highest when we rely on subjective interpretation of objects, where there is no direct access to objects and where there is no formal and extensive representation of a class. High levels of uncertainty may result in higher probabilities of misclassification.

Today we are not likely to encounter uncertainty about an object because the classifier – human or machine – will have the object in hand or will be able to access it in its entirety in digital form. It is more probable that we will encounter uncertainty about a class. While humans have constructed hierarchical classification schemes for centuries, often they have not provided rigorous characterizations of those classes sufficient to reduce uncertainty in the decision process. For example, classification schemes are often represented through narrative scope notes (Figure 2):
through dictionary definitions (Figure 3);
- by default through subclasses (Figure 4);
- by de facto practice as defined in collections (Figure 5); and
- through associated subject headings and descriptors (Figure 6).

In each of these cases, the classifier has little explicit knowledge to work with. As a result, the choice is made based on a subjective interpretation of the class. A human classifier relies on personal subject knowledge and experience. The choice made by the machine classifier will be simple word matching and relevancy ranking.

Optimizing the Classification Decision Using Extensive Class Descriptions

Our first evaluation criterion for a classification decision was the alignment or goodness of fit of an object to the chosen class. The challenge is that we likely don’t have enough information about the class to conduct a good evaluation. Uncertainty rules in this situation [36]. The easiest way to reduce uncertainty is to provide a full and explicit representation of class, its properties and values. Such representation is not a trivial task, though. Today subject experts and human classifiers rely on a deep understanding of a field that they have built up over time. What we need is a way to efficiently and reliably create a full and explicit class definition that can be used to evaluate the choice of class for any object. And, to evaluate that choice, we need an objective, quantifiable and verifiable approach.

FIGURE 4. FAO (Food and Agricultural Organization) Classification of Livestock. Default definition through subclasses [39].

<table>
<thead>
<tr>
<th>Class</th>
<th>Subclass</th>
<th>Class Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Large Ruminants</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Cattle</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Buffalo</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Yaks</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Small Ruminants</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>Sheep</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>Goats</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Pigs or Swine</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Equine</td>
</tr>
<tr>
<td>41</td>
<td></td>
<td>Horses</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>Llamas and Alpacas</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Poultry (Cont.)</td>
</tr>
<tr>
<td>61</td>
<td></td>
<td>Chickens</td>
</tr>
<tr>
<td>62</td>
<td></td>
<td>Ducks</td>
</tr>
<tr>
<td>63</td>
<td></td>
<td>Geese</td>
</tr>
<tr>
<td>64</td>
<td></td>
<td>Turkeys</td>
</tr>
<tr>
<td>65</td>
<td></td>
<td>Guinea Fowls</td>
</tr>
<tr>
<td>66</td>
<td></td>
<td>Pigeons</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td>Other Poultry</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Other Animals</td>
</tr>
<tr>
<td>71</td>
<td></td>
<td>Deer, Elk, Reindeer</td>
</tr>
<tr>
<td>72</td>
<td></td>
<td>Fur-Bearing Animals</td>
</tr>
<tr>
<td>73</td>
<td></td>
<td>Dogs and Cats</td>
</tr>
<tr>
<td>74</td>
<td></td>
<td>Rabbits and Hares</td>
</tr>
<tr>
<td>75</td>
<td></td>
<td>Other (e.g., ostriches, emus, elephants)</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Insects</td>
</tr>
<tr>
<td>81</td>
<td></td>
<td>Bees</td>
</tr>
<tr>
<td>82</td>
<td></td>
<td>Silkworms</td>
</tr>
<tr>
<td>83</td>
<td></td>
<td>Other Worms or Insects</td>
</tr>
</tbody>
</table>

FIGURE 5. Class default definition from library catalog collections.

<table>
<thead>
<tr>
<th>Num Mark</th>
<th>SUBJECTS (1-27 of 27)</th>
<th>Entries</th>
<th>Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Animal Culture -- 13 Related Subjects</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Animal Culture</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Animal Culture America History</td>
<td>2003</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Animal Culture Argentina Periodicals</td>
<td>2004</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Animal Culture Brasil Periodicals</td>
<td>2007</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Animal Culture China</td>
<td>2007</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Animal Culture Classification Books</td>
<td>1992</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Animal Culture Cuba Periodicals</td>
<td>1979</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Animal Culture England Early Works</td>
<td>1800</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>Animal Culture England Early Works</td>
<td>1800</td>
<td>163</td>
</tr>
<tr>
<td>11</td>
<td>Animal Culture England Periodicals</td>
<td>1800</td>
<td>2</td>
</tr>
</tbody>
</table>

FIGURE 6. Class defined through associated subject headings or terms [40].

LIVESTOCK

<table>
<thead>
<tr>
<th>Variants</th>
<th>Narrower Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal husbandry</td>
<td>Cattle</td>
</tr>
<tr>
<td>Farm animals</td>
<td>Donkeys</td>
</tr>
<tr>
<td>Live stock</td>
<td>Draft animals</td>
</tr>
<tr>
<td>Stock (Animals)</td>
<td>Female livestock</td>
</tr>
<tr>
<td>Stock and stock-breeding</td>
<td>Feral livestock</td>
</tr>
<tr>
<td>Broader Terms</td>
<td>Goats</td>
</tr>
<tr>
<td>Agriculture</td>
<td>Hinnies</td>
</tr>
<tr>
<td>Animal culture</td>
<td>Horses</td>
</tr>
<tr>
<td>Animal industry</td>
<td>Male livestock</td>
</tr>
<tr>
<td>Domestic animals</td>
<td>Mini livestock</td>
</tr>
</tbody>
</table>

Special Section
Livestock – domestic animals, such as cattle or horses, raised for home use or for profit, especially on a farm.
One approach that appears to work leverages a combination of machine and human methods. The first step in this process is to assemble a rich but representative sample of objects for the class. It must represent a variety of perspectives — expert-novice, popular-academic, brand specific-generic. And it must represent all aspects or facets of the class. To this collection we apply natural language processing and concept extraction methods to construct a draft representation. Domain experts and classifiers review and revise the draft representation, perhaps several times. When the representation has passed their review, it may serve as an explicit representation of a class.

Figure 7 provides an example of a fully elaborated class representation for Livestock that was generated using this approach. This is one of 750 classes in a scheme. The representation comprises 3,341 concepts that were reviewed and approved by domain experts and professional indexers.

**FIGURE 7.** Elaborated class definition for livestock and animal husbandry — machine classification [41].

Evaluating the Classification Decision

Given an extensive representation of a class, we can make a strong classification decision. It also supports an objective and verifiable evaluation. This approach provides the information we need to evaluate our first criterion — the goodness of fit of an object to a class. Generally, a good fit will result from a high number of matching properties and a high occurrence of those properties. Figure 8 illustrates the way in which a machine categorization engine might report on the goodness of fit to one or more classes. The classification engine can swiftly conduct a property-by-property, value-by-value comparison of the object and class.

**FIGURE 8.** Evaluating the classification process.

- **Sample Object**: Journal article to be classified contains a total of 6,000 noun phrases.
- **Matching Classes**: Livestock and Animal Husbandry
  - **Goodness of Fit**: 1,240 of the 6,000 phrases are found in this class.

Our second evaluation criterion for the classification process pertained to minimizing the risk of misclassification. We can better manage misclassification when we have a full picture of goodness of fit of an object to all classes in the classification scheme. A goodness of fit indicator can be...
calculated for any class where a full representation is available. Figure 9 illustrates the way in which a machine categorization engine might report on the goodness of fit to all classes. Using this approach, institutions may establish thresholds for classification decisions that can be monitored and evaluated. Misclassification is minimized where we can explicitly see which classes may have been overlooked or which may have been selected in error.

Conclusions and Observations

We considered evaluation of a hierarchical taxonomy or classification scheme and the classification decisions made when working with a hierarchical taxonomy. We offer three observations for evaluating hierarchical taxonomies.

1. The principles we need to evaluate and improve classification schemes are readily available. While they are understandable to information science students, some interpretation is needed for designers, engineers and the general public.

2. We can convert these principles into institutional checklists to support periodic evaluation and improvement of classification schemes.

3. Information science education should include assessment of general hierarchical taxonomies in the curriculum, in addition to introducing students to the commonly used classification schemes.

In regard to evaluation of classification decisions, we offer four observations.

1. Classifications are often evaluated indirectly. We have suggested an approach that more directly targets the classification decision directly.

2. This approach requires more information about the classes in a classification scheme. It makes explicit the implicit knowledge of classifiers used to make decisions. Providing more information about a class is not a trivial task for existing schemes. However, it is manageable for new schemes.

3. The approach allows an institution to objectively judge the goodness-of-fit of any decision and to assess the risk of misclassification.

4. This approach supports rather than substitutes for other evaluation perspectives. Understanding the nature of the classification decision helps us to better understand end user responses to that decision.

While we considered evaluation of the scheme and the decision separately, we should not overlook the dependencies between a well-formed classification scheme and a well-executed classification decision. Making an optimal classification choice is dependent upon a good representation of the class, a well-formed class and a well-formed classification scheme. While the context in which taxonomies are used has expanded significantly in the past 20 years, the criteria for evaluation have not changed. The expansion in affordable computing power and the availability of semantic technologies provides the capacity to make and evaluate classification decisions in low risk and object ways.
Resources Mentioned in the Article


Resources continued on next page
### Resources Mentioned in the Article, cont.

<table>
<thead>
<tr>
<th>Number</th>
<th>Reference</th>
</tr>
</thead>
</table>
Case Study: Developing the PLOS Thesaurus
by Jonas Dupuich and Gabe Carr

When a good thesaurus is working well, you barely notice it. People searching your website click on relevant terms and find relevant results. Those responsible for creating metadata do so with pleasure and ease. “Yes,” they may say, “this concept perfectly captures a key aspect of this work.”

Thesauri less suited to their tasks lead to any number of difficulties. Creating metadata becomes a slow and difficult task, and results suffer. Discovery is hindered, and users shy away from relying on selected terms. Analysis becomes impossible.

In late 2011, the Public Library of Science (PLOS) found itself struggling to get value from its thesaurus. Key terms were missing across a variety of disciplines while over 1000 terms went unused. It was time for a change.

To gain a better understanding of the problem, PLOS contracted with Access Innovations (AI), an information services provider, to analyze its thesaurus. AI created a study with the aims of

- creating a robust and standards (Z39.19) compliant version of the thesaurus adequate for machine analysis;
- testing the thesaurus against the published PLOS corpus – over 21,000 papers;
- identifying the work needed to produce a thesaurus that meets PLOS needs; and
- developing a budget for the work.

AI completed the study in January 2012. It identified over-used terms, under-used terms and unused terms. Radial graphs highlighted imbalances among term distribution. Most importantly, it provided a path forward. Key recommendations provided general guidance about term selection:

Jonas Dupuich is product manager at the Public Library of Science (PLOS). He can be reached at jdupuich@plos.org.

Gabe Carr is a taxonomist at Access Innovations. He can be reached at gabe_carr@accessinn.com.
All terms should be ones that would normally be typed into search queries.

All terms should represent concepts in articles covered by PLOS journals.

In addition, it provided specific guidance about how to build out the PLOS thesaurus:

- Review as possible synonyms, the terms used less than five times.
- Break up terms used more than 1,000 times into smaller conceptual areas.
- Review unused terms with an eye toward removal.
- Check for under-indexed articles as a way of analyzing for missing concepts.
- Add terms from AI’s STEM thesaurus.

One of the biggest outstanding questions for PLOS was whether existing thesauri would meet its needs or whether rebuilding its current thesaurus would provide the most effective starting point. AI’s study recommended the latter.

Well-known thesauri like MeSH, the National Library of Medicine’s controlled vocabulary thesaurus, are great for the areas they cover, but they tend to be deep within defined fields or shallow across a broad array of subjects. As a multi-disciplinary publisher, PLOS publishes papers that don’t fit into neat categories. For a thesaurus to successfully represent content across all biomedical fields as well as disciplines as diverse as information science and paleontology, PLOS needed a custom-built thesaurus.

With a clear understanding of the project scope and the resources required to complete it, PLOS committed to rebuilding the thesaurus.

**Rebuilding the Thesaurus**

Before starting the work to rebuild the thesaurus, PLOS collected the use cases that would define its utility and guide its development.

The collection of use cases extended beyond current cases to include future uses for the thesaurus. One of the more painful consequences of an ill-suited thesaurus was that it became a barrier to the adoption of new services designed to leverage well-classified content.

As a primary use case, PLOS uses terms from its thesaurus to enhance the metadata of its published papers. Once terms have been assigned to a given paper, they assist in discovery services, including the search and browse features available on PLOS websites. Future cases include strategic activities like trend and data analysis and identifying gaps in editorial boards. Table 1 illustrates a complete list of use cases.

A final decision leading up to the rebuild was vendor selection. As AI had outlined the scope of the project and completed a lot of the prefatory work, PLOS was able to leverage this effort by contracting with AI to rebuild the thesaurus.

To kick things off, PLOS delivered the following to AI:

- current copy of the PLOS thesaurus
- use cases for the thesaurus
- over 2,000 changes and additions collected by editorial staff
- guidance for developing a research analysis and methods branch

**TABLE 1. Current and proposed use cases for the PLOS thesaurus.**

<table>
<thead>
<tr>
<th>Discovery</th>
<th>Workflows and Matching Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Connecting users to content via search and browse services on PLOS websites (current)</td>
<td>- Suggesting academic editors for submissions (current)</td>
</tr>
<tr>
<td>- Generating email research alerts (current)</td>
<td>- Suggesting reviewers for peer review (current)</td>
</tr>
<tr>
<td>- Generating RSS feeds (current)</td>
<td>- Providing targeted instructions to authors, editors and reviewers (proposed)</td>
</tr>
<tr>
<td>- Recommending related content (proposed)</td>
<td>- Identifying related papers for editors and reviewers during peer-review (proposed)</td>
</tr>
<tr>
<td>- Creating content collections (proposed)</td>
<td>- Connecting people to articles or parts of articles such as tables and figures (proposed)</td>
</tr>
<tr>
<td>- Enriching metadata of non-article content associated with the article (proposed)</td>
<td>- Connecting people to PLOS products or services (proposed)</td>
</tr>
<tr>
<td>- Generating meta-tags used by services like Google Scholar (proposed)</td>
<td>- Matching papers to staff, section editors, etc. (proposed)</td>
</tr>
<tr>
<td>- Generating microformats (proposed)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategic Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Trend and data analysis by taxonomic classification (proposed)</td>
</tr>
<tr>
<td>- Identifying gaps in PLOS editorial boards (proposed)</td>
</tr>
<tr>
<td>- Targeted advertising (proposed)</td>
</tr>
</tbody>
</table>
As a resource for expanding the thesaurus, AI drew upon an internally created science, technology, engineering and medicine thesaurus (STEM). AI produced a list of approximately 15,000 STEM terms that were absent from the existing thesaurus and the frequencies with which they appeared in the PLOS corpus. From this list, AI added those terms that were judged to be sufficiently significant and specific enough to be valuable for detailed subject coverage and optimum indexing accuracy.

During the rebuild process, the structure of the hierarchy developed naturally, with the placement of new terms often suggesting the need for new branches or reorganization of existing ones. PLOS met with AI weekly to discuss any concerns that arose during the week. These conversations usually revolved around basic questions of term organization such as, “Where does cognitive science belong?” “What is the relationship between information science and computer science?” “Will we limit the thesaurus to five tiers or six – or more?”

Following AI’s progress over the course of the project was easy as AI provided PLOS with web access to the current draft thesaurus. While AI put the finishing touches on the thesaurus, PLOS lined up reviewers and subject matter experts to provide feedback on AI’s draft. This allowed PLOS to begin the review the day after the draft arrived.

The review focused on two tasks – ensuring the right terms were included in the thesaurus and ensuring the terms were organized in a fashion that suited the PLOS use cases. The latter task proved difficult. On one hand, PLOS wanted the terms to be organized in an absolute sense that accurately reflected the relationships between terms. On the other hand, PLOS wanted the terms to provide users with clear categories that represented the areas in which it publishes. The tension between these goals extended the review period beyond the original estimate of two weeks, but it ultimately allowed PLOS to address important questions that were not identified up front. An outline of the review process included the following steps:

- Senior editors provided feedback about the top-level terms.
- A subject matter expert used this feedback to create a candidate list of top level terms.
- A thesaurus policy group reviewed all feedback and forwarded its recommendations to AI.

- AI suggested minor enhancements, which were later approved by the thesaurus policy group.

The result was a 10,400-term thesaurus comprising seven tiers. Table 2 displays a list of top-tier terms.

### Applying the Thesaurus

Rebuilding the PLOS thesaurus goes a long way toward addressing PLOS needs for a controlled vocabulary, but it only gets at half the problem. In addition to identifying deficiencies in the PLOS taxonomy, AI’s initial study made it clear that terms had not been adequately applied to the published corpus. Typically, submitting authors select terms from the PLOS taxonomy to identify concepts addressed in their papers. Journal staff supplement these terms to ensure that submissions have a minimum number of appropriate terms.

While this process captures many of the more important concepts present in papers, significant numbers of concepts were being neglected. This finding is not surprising given the approach. Machine suggestions can be far more complete than manual suggestions. Experts selecting from machine-suggested terms tend to produce even better results, but as this approach isn’t currently available, PLOS decided to investigate machine-aided indexing to improve its subject area metadata.

Relying on industry contacts and previous experience, PLOS identified candidate software applications to perform machine-aided indexing. After analyzing the top contenders, PLOS selected AI’s MAIstro for the job.

One of the benefits to PLOS of the MAIstro approach is that indexing relies on a flexible rule-based system that can be enhanced over time to produce high quality results. By default, every thesaurus term has a simple rule (for example, apply “Biology” when the term biology appears in the text). By applying conditions, such as requiring upper or lower case, proximity with other text strings or position within the sentence, AI creates complex rules that can better isolate concepts and index with more appropriate terms.

<table>
<thead>
<tr>
<th>TABLE 2. Top-tier terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Biology and life sciences</td>
</tr>
<tr>
<td>2. Computer and information sciences</td>
</tr>
<tr>
<td>3. Earth sciences</td>
</tr>
<tr>
<td>4. Engineering and technology</td>
</tr>
<tr>
<td>5. Environmental sciences and ecology</td>
</tr>
<tr>
<td>6. Medicine and health sciences</td>
</tr>
<tr>
<td>7. Physical sciences</td>
</tr>
<tr>
<td>8. Research and analysis methods</td>
</tr>
<tr>
<td>9. Science policy</td>
</tr>
<tr>
<td>10. Social sciences</td>
</tr>
</tbody>
</table>
The better the rule base, the better the metadata. At AI’s suggestion, PLOS will aim for a mix of 80% simple rules and 20% complex rules as a starting point before integrating the service into its production workflow.

The rule building is proceeding as this article is being written. Upon receipt of the final rules, PLOS will test sample output, make necessary adjustments and prepare for implementation. Implementation will include the following steps:

- PLOS will submit article text to MAIstro.
- MAIstro will return terms and term frequencies to be indexed for search purposes.
- PLOS will re-index all previously published articles (now over 50K articles).
- PLOS will make the new terms available to users in Editorial Manager, the peer-review workflow management system produced by Aries Systems.

As AI is working to complete the rule building, PLOS is preparing an implementation plan. As a starting point, PLOS is looking to submit the following sections of published content to MAIstro for indexing:

- Title
- Abstract
- Methods and Materials (when present)
- Results (when present)

From these sections, PLOS will add the top seven terms returned by MAIstro in an effort to capture the key concepts present in submitted articles.

Although PLOS doesn’t expect to obtain optimal results from this approach as trade-offs are involved, it believes these additions will be a good starting point from which it can learn and continue to improve in the future.
Building User Experiences: Synchronizing User Experience Design and the Supporting Metadata and Taxonomy Infrastructure

by Carol A. Hert, Gary Carlson and Bram Wessel

EDITOR’S SUMMARY
For best results in website architecture and operation, the process of user experience design and taxonomy and metadata development should be synchronized, not developed in isolation. Bringing the two work streams together enables full consideration of how users, represented as personas, would interact with the site, pursuing likely needs and preferences. The persona review informs the technical infrastructure and functionality as well as the terminology and content types. Analysis of the site design and comparison with similar sites can lead to better appreciation of the user experience and expanded vocabulary links to improve access to content. The Key Paths methodology used by the design firm Factor captures likely user journeys to define design elements, allowing for multiple routes to appropriate content. Metadata must be available to support design, and functional design and content tagging must take advantage of metadata. The dynamic interaction and mutual support of the site’s content team and design team are keys to an effective product.

KEYWORDS
user experience  user models  metadata
information architecture  taxonomies

Despite their best intentions, user experience designers and taxonomy and metadata developers have often found that their work is not well connected, even though both are highly interrelated. For example, a design might be proposed that needs segmentation of content by user role, but there may not be metadata associated with content that captures the role, resulting in the need for detailed review of content and hand coding to create the experience. Taxonomists might build a taxonomy for roles without knowing which roles the design uses, leading to over- or under-specification of the taxonomy. In our recent projects, Gary Carlson Consulting and Factor have been tightly integrating these two different work streams, assuring that the experience design and metadata and taxonomy infrastructure and their governance and maintenance are synchronized and integrated (Figure 1).

Why bring these two different work streams into conjunction? Gary Carlson Consulting (GCC) is often asked to provide a metadata and taxonomy strategy development. Essentially, the job is to specify the information model (metadata components, taxonomies and business rules associated with those components) and its governance to support
business goals, including enhanced user experiences for search, online shopping and other tasks. Meeting these business goals often requires a coherent enterprise-scale content strategy that depends heavily on this information model. Many user experience projects are also driven by these exact same goals. Factor’s human-centered experience design practice drives towards building experiences that, along with being driven by user goals, are implementable and maintainable without heroic efforts; that is, they rely on metadata and taxonomic structures to guide navigation as well as assure appropriate content provision. For example, we might recommend that a website provide a “Company in the News” component that is populated automatically via usage of metadata about content (in this case, date of creation and a content type of news release).

GCC’s work is dependent on knowledge of the design – what is the organization attempting to accomplish via what experiences? What navigational structures and key pathways are proposed? What stakeholders need to engage with and maintain the model? Factor’s design work must present content in relevant and context-sensitive user experiences. To develop the most effective design Factor relies on a shared understanding of information modeling considerations such as how content can be tagged, what content types are in use and what structured data can be incorporated into the experience.

Both teams have found that working in isolation from each other means that relevant information does not get shared at the right time, the technical content strategy can be divorced from the end design and the end design either doesn’t fully exploit available metadata or is reliant on information about content and users that is not easily inferred. The result of not integrating these two practices is often that a project may meet the defined goals in the short term, but that the long-term success is placed in jeopardy due to a lack of coordination of the user design and the information model. How does our approach work in practice? We’ll look at several examples of typical work products used by both teams to indicate the connections.

**Example 1: Informing Technical Infrastructure through Persona Review**

Personas represent the “human” in human-centered design. They are the means through which user goals, mental models and information needs are captured in believable narratives and expressed in effective experience designs. The technical content strategy team can also use personas to inform types of taxonomies to build and to identify metadata elements. Figure 2 shows a persona for a commercial site selling nutritional and health products without taxonomy.

The technical content strategy team’s review extracts potential taxonomies, content types and functionality that would support this persona’s experience. In Figure 3 above, we highlight insights from the persona that directly inform the information model.
Example 2: Design Informs Taxonomy Development

As part of a recent web project, a taxonomy of audience types was needed for an educational institution that had a large and diverse undergraduate population. The technical content strategy team initiated the work stream by reviewing the existing website along with sites of comparable institutions and by conducting research interviews with the client project team. As the design team began developing personas, defining the design and modeling the user experience, the aforementioned research and analysis suggested that the undergraduate population was quite diverse and needed multiple entry points and paths through the content. This, in turn, led to further specification in the taxonomy to assure tagged content could support the experience as shown in Table 1.

Example 3: Extracting Metadata Requirements from Key Paths

A methodology used by Factor’s design team is to model user journeys through content in order to articulate design components. The resulting key paths are critical to the technical content strategy team to confirm that the information model includes the right content types, taxonomies and related functionality to segment content for automatic provision. Below is a snippet of a key path modeling the journey of a medical patient who has some symptoms of a condition and wants to understand treatment options. The path indicates several possible entry points and types of content/experiences to be provided. From the technical content strategy team’s perspective, the flow provides further insight into needed content types: articles on conditions, discussion groups and potentially a directory of doctors or an online form to input conditions, symptoms or treatments to find doctors. Additionally, the path indicates that taxonomies for conditions, symptoms and likely treatments will be needed and that they need to be related to each other and to doctors that specialize in those areas.

These examples capture the synergistic relationship of the design and technical content strategy teams throughout the integrated process. The technical content strategy team determines the answers to questions such as:

- Can the information model support the information delivery specified by the design?
  - Do all the taxonomies exist and are they up-to-date?
  - Do the taxonomies support the design experience?
  - When taxonomies are exposed in the design (for example, in navigational aids) is terminology appropriate for the users?

### TABLE 1. Revisions to the taxonomy of audience types

<table>
<thead>
<tr>
<th>Initial Draft Based on Comparative Website Reviews and Stakeholder Discussions</th>
<th>Final Version After Ongoing Engagement with Design Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrators</td>
<td>Administrators</td>
</tr>
<tr>
<td>Alumni</td>
<td>Alumni</td>
</tr>
<tr>
<td>Community Neighbors</td>
<td>Community Neighbors</td>
</tr>
<tr>
<td>Faculty and Staff</td>
<td>Faculty and Staff</td>
</tr>
<tr>
<td>Parents</td>
<td>Graduate Students</td>
</tr>
<tr>
<td>Prospective Students</td>
<td>Parents</td>
</tr>
<tr>
<td>Students</td>
<td>Prospective Students</td>
</tr>
<tr>
<td>Students New to the University</td>
<td>Undergraduate Students</td>
</tr>
<tr>
<td>Full-time Students</td>
<td>Students New to the University</td>
</tr>
<tr>
<td>Graduating Students</td>
<td>Full-time Students</td>
</tr>
<tr>
<td>Commuting Students</td>
<td>Graduating Students</td>
</tr>
<tr>
<td>Part-time Students</td>
<td>Commuting Students</td>
</tr>
<tr>
<td>Note that student types have expanded representing different paths for students in the site; content can be tagged with multiple student terms reducing the problems associated with non-mutually exclusive terms in this area.</td>
<td>Note that student types have expanded representing different paths for students in the site; content can be tagged with multiple student terms reducing the problems associated with non-mutually exclusive terms in this area.</td>
</tr>
</tbody>
</table>
Can the desired relationships between taxonomies and content be modeled?
- Are the relationships well enough defined to be applied to the concepts?
- Do we have the business rules needed to drive content to for the experience?

The design term is better able to answer questions such as

- Can the design be executed effectively and efficiently via available metadata?
- Is the content tagged and managed such that it can be implemented in the design?
  - Are all the required taxonomies in place and up-to-date?
  - Do the tags support the design goals?

- Are our templates for content creation synched to the content types?
- Do we have access to the correct metadata to support search, navigation and other design aspects?
  - Does the correct information model exist?
  - Does the technical infrastructure allow search and navigation functionality to access the information model in a viable and scalable manner?

- Are the systems using the information model synchronized appropriately?

These examples provide a very brief sense of some of our points of integration. We also work jointly on ongoing maintenance, including strategy and implementation of editorial and information model governance. Again, these two need to be synchronized to support our customer’s needs for appropriate and sustainable designs.

We have found that by tightly integrating our efforts both teams and clients benefit. The experience design team builds designs that are sustainable and manageable using available metadata, and the technical content strategy team has full access to the design considerations that drive its work. The customer gets a design that is resonant with the organization’s goals and implementable over both the short and long term of the project.

Acknowledgement
The ideas referenced in this article benefited in part from work conducted by the authors under a brand-licensing agreement with The FatDux Group.
Indexing Electronic Medical Records Using a Taxonomy
by John Kuranz and Barbara Gilles

EDITOR’S SUMMARY
With the move toward patient electronic medical records (EMRs), accessing information for insurance coding and research depends on standardized taxonomies to organize and index the content. Controlled vocabularies are necessary to interpret content consistently. Established quasi-taxonomies provide codes for medical conditions and treatments, but applying these codes as metadata to index the records is laborious, requiring translation from natural language in the EMR to a code’s verbal equivalent to the code. Indexing systems can streamline the categorization process for greater efficiency and accuracy by using Bayesian engines or a rule-based approach. Analyzing discrepancies between human indexing and the software system results shows where editorial intervention is needed for continual improvement, with a goal of 85% or higher accuracy. Using a categorization system with a hierarchical taxonomy enables deep, precise indexing or quick, automatic filtering to more general concepts. The accuracy of medical indexing systems varies widely, based on the degree of automation and capacity for semantic analysis.

KEYWORDS
medical records  knowledge bases
subject indexing  semantic analysis
machine aided indexing  taxonomies

John Kuranz is CEO, Access Integrity. He can be reached through the company website: www.accessintegrity.com.
Barbara Gilles is a taxonomist at Access Innovations, Inc. She can be reached at barbara_gilles<at>accessinn.com.

The use of electronic medical records (EMRs) has increased dramatically over the past few years. In the health industry and in medical research, EMRs have proven to be a valuable source of information for diagnostic research, coding enhancement for billing and insurance purposes and understanding of an overall patient encounter. Studies to date have revealed that the text content in EMRs may contain important additional information relevant to outcomes, concomitant diseases, procedures, interventions or test results in observational studies. However, completely manual review of EMRs is time-consuming, subjective and inefficient. Furthermore, manually applying codes and/or terms as metadata to index records is laborious. The solution to these problems is the use of machine-aided indexing software to extract diagnostic and other clinical information.

Because EMR text can contain a wide range of complex language structures, terminology and context-specific abbreviations and acronyms, the software must be able to handle those aspects of clinical language. To provide a foundation for such complex language analysis, an organization needs to select or develop an appropriate taxonomy. This provides a consistent vocabulary for making the indexing consistent. At the same time, it provides a basis for rule building (in the case of rule-based indexing systems) or system training (in the case of Bayesian indexing systems).

Building or Borrowing a Taxonomy for EMR Indexing
Consider what, essentially, you want to cover in your indexing, and make sure you cover it in your taxonomy. Diagnoses, and therefore diseases, injuries, conditions and symptoms? Surgical procedures, medications and other treatments? There are coding systems and taxonomies that cover these areas.
If you are creating a taxonomy from scratch, enlist the help of subject matter experts (SMEs) who are familiar with health services and medical terminology and have them suggest and review terms. You might need some generalists to spot ambiguous and overlapping terms among the various disciplines and sub-disciplines. The SMEs should also review the hierarchical structure after it has been roughed out.

Due to the needs and expectations of the medical community, your medical taxonomy will need to be very specific. From what we’ve observed, the term hierarchy will need to go at least five or six levels deep. Depending on the services (primary care, for example) and specialties (cardiology, for example) your organization covers, the overall scope might be fairly narrow or all-encompassing. For a narrower scope, you might need only 10 top terms.

One of the standard coding systems can serve as the basis for the taxonomy. The 9th edition of the International Statistical Classification of Diseases and Related Health Problems (ICD-9) and the American Medical Association’s Current Medical Terminology (CPT), among others, are large coding systems heavily used by the medical provider community. However, these coding systems are not taxonomies per se. And this may be an advantage for associating required codes with EMRs. However, it means that things start out somewhat backwards. For “indexing” a code, it is logical and practical to leave it as a taxonomy “term,” rather than change it to a descriptive word or phrase. However, these terms are not easy for a taxonomist, automated rule builder or human indexing rule editor to deal with without some reference to semantic content. One way of adding the semantic content back in is to put it in a synonym or non-preferred term field. Figure 1 shows an example from the ICD-9 diagnosis code set, with the descriptive phrase in the synonym field coming from the ICD-9 tabular list that explains the codes.

Software Training and Rule Building

In a rule-based indexing system, we can easily view the synonym and write and edit rules to capture the various ways in which the concept above (“food poisoning due to Vibrio vulnificus”) might be expressed in an EMR (Figure 2).

Unfortunately, the descriptions in the medical codebooks are not always as neat and clear as in this example. Some very common problem descriptions are “Other,” “NOS” (not otherwise specified), and “NEC” (not elsewhere classified. And then there are the multiple occurrences of descriptions (such as “varicella”) that have different meanings, depending on what section they’re in.

Good taxonomists and rule builders will check the context before building the rules that differentiate among the codes for currently having a disease, having a history of the disease, having a recent exposure to the
disease and needing to be vaccinated against the disease (Figure 3). You’ll still need the codebook for reference.

An indexing system needs to be trained in the specific subject or vertical concept area. In rule-based systems, this training is accomplished by (1) selecting the approved list of keywords to be used and, through matching and synonyms, building simple rules; and (2) employing phraseological, grammatical, syntactical, semantic, usage, proximity, location, capitalization and other algorithms – based on the system – for building complex rules. This approach means that, frequently, the rules are keyword-matched to synonyms or to word combinations using Boolean statements in order to capture the appropriate indexing terms in the target text.

In Bayesian engines, the system begins with a training set of indexed patient records, usually 50-60 documents. The system uses the sample to associate the keywords with text, creating scenarios for word occurrence based on the words in the training documents and how often they occur in conjunction with the approved keywords for that item. Some systems use a combination of Boolean and Bayesian engines to achieve the final indexing results.

In the world of knowledge management, the measure of the accuracy of an indexing system is based on the number of hits (exact matches with what a human indexer would have applied to the system); misses (the keywords a human would have selected that a computerized system did not); and noise (keywords selected by the computer that a human would not have selected). The statistical ratios of hits, misses and noise are the measure of how good the system is. Our experience has been that, for a system to be practical and beneficial, the threshold should be at 85% hits out of a total of 100% accurate (against human) intervention. That means that noise and misses need to be less than 15% combined.

A good system will provide an accuracy rate of 60% initially from a good foundation keyword list with simple match rules and 85% or better with training or rule building. This low starting point means that there is still a margin of error expected and that the system needs – and improves
with – human review. Some systems can maintain a statistical record of discrepancies between human indexing and the automated term suggestions or assignments; ideally, the sampling and review that this approach requires should be done at least once every few months.

While human monitoring and control of the indexing of individual records is ideal, perceived economic or workflow impacts often render frequent human participation in the indexing process unacceptable. These issues generally lead to the attempt to provide some form of fully automated indexing results, so that human indexing is not required. Fortunately, there are some techniques, other than rule development or system training, that can help increase accuracy:

- The system can be configured in such a way that only the most specific terms are used; this practice prevents the application of terms that might be considered too general for the text. (For coding, though, the rules should be written so that only the most specific code numbers are used.)
- If coding is not involved, the keywords may also be “rolled up” to ever-broader terms until only the first three levels of the hierarchy are used; this procedure casts a broader net and prevents the use of terms that might be considered too specific for the context. This second approach is preferred in some environments, where popular thinking indicates that users will not go deeper into the hierarchy. Deeper indexing and precise application of keywords still benefit from human intervention, at least by review, in all systems. The decision then becomes how precisely and deeply the user develops the indexing for the system application and the target user groups.

**Summary**

We have investigated some methodologies used in the automatic and semi-automatic classification of text in the medical field. In practice, many of the systems use a mixture of the methods to achieve the result desired. Most systems require a taxonomy in order to start, and most systems tag text to each keyword term in the taxonomy as metadata in the keyword name or in other elements. A taxonomy that has been so enhanced enables deep, precise indexing or quick and automatic filtering to more general concepts.

There are real and reasonable differences in deciding how a literal world of data, knowledge or content should be organized. Purveyors of various systems maneuver to occupy or invent the standards high ground and to capture the attention of the marketplace, but they often bring ambiguity to the discussion of process and confusion to the debate over performance.

The processes are complex and performance claims require scrutiny against an equal standard. Part of the grand mission of rendering order out of chaos is to bring clarity and precision to the language of our deliberations. In simple terms, it’s about how to bridge the distance between questions from humans and answers from systems. When the answers are in your EMRs, taxonomies and associated indexing systems can bridge the gap.
Welcome to the second decade of online taxonomy construction and maintenance!

In the early days of online taxonomies, our focus was primarily on navigation. We were using online taxonomies to improve the browsing experience on websites, and occasionally, when good metadata was available, we used taxonomies to improve search. But we’ve come a long way since then, and our focus has changed. We now build taxonomies to support auto-classification systems.

The early efforts are exemplified by the categories you see on e-commerce websites like eBay or Amazon; these sites include high-level, general-to-specific categories that allow users to drill down, or navigate, to exactly the product they need.

When we started creating online taxonomies, auto-classification was in its infancy, and the software was not really reliable yet. There was a huge disconnect between the high-level structure of a navigational taxonomy and the data requirements of auto-classification software.

More often than not, early auto-classification software seemed to be overly ambitious. Many early attempts involved clustering techniques that were used to collect documents that included similar “bags of words.” The most unique words often occurred at the very specific detail level, so the clusters tended not to make sense on their own. They required a suitably descriptive, higher-level label.

But technology has evolved, and we are in a different situation today. With the advent of sophisticated content analytics techniques, entity extraction is increasingly reliable, and we now routinely use proper noun entities, as well as noun phrases, to enhance our classification schemes.

Taxonomy building has also changed as a result of these advances in entity...
and phrase extraction techniques. When we build taxonomies now, we still work to create descriptive high-level categories, but we also work to accurately associate the lower-level entities with the higher-level categories we’ve created. We provide an overview of this taxonomy building process here, although each step in the process really deserves an article in its own right.

**Entities and Text Mining**

Entities are phrases that identify people, places and things that are easily recognized by content analytics software. Entities are often recognized by specific patterns in text. Place names are good examples of entities, but entities can also be based on lists of terms that already exist, such as programming languages (Figure 1).

Entities often supply the best evidence for identifying higher-level taxonomy tags, and text-mining tools are now very adept at identifying entities in content. You’ll likely use text-mining tools at varying points in your taxonomy development cycle.

For example, you might want to use a text miner to identify product names in your content. In this case, your highest-level taxonomy category would be Product, and you would likely also have a metadata tag, possibly called “MyCompany’sProducts,” that is associated with your content. The values for the “MyCompany’sProducts” tag are drawn from the terms in the Product taxonomy. You might also include second level categories that describe product types.

If your taxonomy supports a high-tech site, your secondary tags could be Software Products, or Open Source Software, followed by specific product names, like Lucene. You probably wouldn’t include the high-level terms as actual values in your metadata tags, since they are too general, but you would include them in your taxonomy so that you can identify what kind of thing the entity is.

Most text-mining tools provide information about how many of each type of entity they find. If, for example, you run a subset of your content through a text-mining tool and see many specific product names identified, you can assume that Product will be a good high-level descriptive category.

Text miners can often identify phrases, and these phrases are useful during the early taxonomy design. If you see many phrases that describe specific activities around a subject area, for example, "programming, software engineering,"
coding or software testing in your content, then “Software Development” is a good candidate value for a Subject or Topic tag. You’ll also use these phrases later, as synonyms, so the auto-classification tools pick up term variants in your content.

**A Basic Methodology**
No matter which tool you use, you should follow this basic methodology when building your taxonomy.

1. Understand the auto-classification technique that your tool uses.
2. Create a taxonomy model.
3. Test your classification.

Your final step describes the way your classification will be used, and this step comes down to cases. If you use your classified content to drive a faceted search user interface (UI), for example, you’d add an additional step to test your classification in the context of the new UI.

**Understanding Auto-categorization Techniques**
Understanding how the various auto-classification products work is key to building an effective taxonomy for categorization. Most focus on one technique, with secondary support for the others, and the best products provide a hybrid approach.

Auto-classification products rely on a good classification model. The model includes the high-level categories, or tags, that will be associated with content and also defines the evidence, or rules, that determine when these tags will be applied. The tools for collecting and analyzing the evidence differentiate the products.

You can often judge the strength of any vendor’s approach by looking at the richness of its taxonomy management offering. Look at the tools that are provided to help you create your classification and also at the tools that are provided to help you test your taxonomy. The tools are essential, but they vary widely by vendor, and robust test tools often indicate a more robust product.

The tools vary in terms of the amount of maintenance required to create an accurate classification. All of the techniques require a good amount of up-front design work and ongoing testing and maintenance. No auto-classification tool is perfect, but software classification is always more consistent than human categorizers.

**Linguistic/Lexical Tools.** Tools that concentrate primarily on a lexical approach will have the richest taxonomy management functionality. Lexical tools let you gather and rank representative words and phrases that are associated with the concepts to be classified; these tools allow you to identify the keywords and phrases the way they occur within the text.

You rely heavily on identifying synonyms and term variants when you use the lexical tools, since you want to identify the various ways that important concepts appear in your content. Often you don’t want to identify only exact phrases, but also words that appear in the same sentence. Products often provide settings that help.

Lexical tools require a great deal of initial up-front analysis and a very iterative development and testing cycle. You will add terms, synonyms and new rules as you notice different occurrences of the important phrases in your content.

**Rules-based Tools.** Rules-based tools provide a rich syntax that you use to control the way the evidence in the taxonomy will be used to add tags to text. As with lexical tools, explicit rules can be used to indicate that you want words to appear in the same sentence, rather than only as a phrase.
But rules are often used to identify words and phrases that should not be used as evidence. Rules-based systems also allow you to compare more than one lexical construct to provide an additional level of control over the classification. They can be used to tell the difference between, or disambiguate, similar terms or entities that appear in your taxonomy. For example, there is a Will Smith who plays American football for the New Orleans Saints, and there is also Will Smith, the actor. If the entity “Will Smith” appears in a sentence, the evidence would indicate that the content could be categorized as either Entertainment or Sports. But if the title *Men in Black* also appears in the same sentence, the sentence is more likely talking about Will Smith the actor and should be classified as Entertainment. Rules allow you to compare the entities and make the appropriate tagging determination.

Right now, there is no common syntax for developing rules; the syntax varies from tool to tool. Rules syntax ranges from the kind of Boolean syntax that is usually associated with search engines to the more complex syntax more commonly used in programming languages. Because of this lack of consistency, the people who create and maintain these rules will have a more specialized skill set and will require more training.

**Machine Learning and Predictive Analysis Techniques.** Machine learning and predictive analysis techniques are interesting because they often initially appear to be less manual-labor-intensive than the other approaches. Although many of these systems claim that they are not using taxonomies on the back end, they are still identifying and classifying phrases and entities. These tools take advantage of the same content analysis and text-mining techniques that taxonomists use to enrich classification schemes, but their models are more complex than a traditional taxonomy structure. And these systems rely on iteration to continuously validate the models they use. Again, you may not create a traditional hierarchical taxonomy to support machine-learning systems, but you will create taxonomies that these systems can use for reference. You might also select representative document sets to train these systems.

Maintenance of machine learning systems always involves repeated training, especially when you add new content. You will also help revise the larger machine-learning model as you learn more about your content.

**Create a Taxonomy Model**

Designing a model for auto-classification is similar to but not exactly the same as designing a taxonomy for manual indexing or navigation. Like more traditional taxonomy construction, both top-down and bottom-up analysis is required, and knowledge of the content to be classified is key.

Typically, you figure out how you want to organize your taxonomy, and then you build it out. You do this task by deciding on the general categories first and then adding more specific terms within each category later. Your model can be hierarchical, and, as we said above, you typically associate the highest-level terms with the actual metadata tag names.

You’ll want to figure out the level of specificity of the terms to include in your tags. Do you want to tag major mentions only, or do you need to include every occurrence of a Product name? Again, there might be many terms in your taxonomy that are used as evidence and not as actual values that appear in your metadata tags.

You’ll often import terms from existing sources or use an existing taxonomy as a basis for your model for auto-classification. But while taxonomies that are used for manual indexing or for navigation can usually accommodate poly-hierarchy, your auto-classification taxonomy model should avoid it. Using the same term as evidence for two separate categories will lower the accuracy in each.
Test Your Classification

Testing your classification is the most important step in the process. Language is imprecise, no matter what subject area your taxonomy covers, and you will be surprised (and often amused) by some of the mistakes an automatic tool can make. You will want to use the test tool to look for inaccurate tags, but you will also want to look for tag omissions.

**Collect a Document Set.** It’s useful to collect a set of documents that you use over and over as you test. If you use the same document set for each test, you can easily track the impact of any changes you make. Since you eventually add content to this set, you can re-test and see how your classification performs as your content changes.

**Configure Your Auto-classification Tool.** Your configuration will depend on the tool you use, and it will usually require some support from your IT department, but you will want to have a basic knowledge of the default rules it will use to do its job. If your tool is basing its classification decisions on words that occur within sentence boundaries, for example, you’ll want to be sure that your content includes sentence boundaries that the tool will understand.

Iteration is key to the testing process, since minor changes in your taxonomy often have major impacts on your classification. No taxonomy is ever really complete, but you will want to understand when your classification is ready to be put into production.

So again, welcome to the brave new world of taxonomy development for auto-classification. While this new methodology has a bit more in common with software development than it does with traditional library science, the outcome is much more satisfying, and you will put your information management skills to good, practical use.
Building Controlled Vocabularies for Metadata Harmonization
by Marcie Zaharee

EDITOR’S SUMMARY
Organizations using controlled vocabularies are highly diverse, but they all rely on using a terminology list that is internally standardized, commonly understood and widely applied for metatagging. The basic steps for creating a controlled vocabulary, taxonomy or thesaurus are the same for each case, starting with determining the scope to be covered and identifying representative content sources. This step is followed by gathering and organizing terms, enhancing them with synonyms and relationships and often by subject-matter-expert review. Using a dedicated taxonomy management tool is recommended to facilitate management, visualization and export in machine-readable format such as OWL or SKOS. Posting the controlled vocabulary to a registry or data warehouse enables sharing and may stimulate broad acceptance. The MITRE Corporation has followed this process in developing taxonomies consistent with the Department of Defense’s Net-Centric Data Strategy, which bases shared understanding within and among DoD programs on the use of controlled vocabularies and crosswalking equivalent data elements.

KEYWORDS
controlled vocabularies information mapping
index language construction standardization
metadata

Marcie Zaharee currently works in the MITRE Center for Integrated Intelligence Systems (CIIS) where she is the lead for developing a controlled vocabulary, taxonomy and thesaurus for the intelligence reconnaissance and surveillance (ISR) community. Marcie joined MITRE in 2005 as the associate department head for information management and practice where she worked to advance knowledge management in MITRE, championing efforts that support staff collaboration, knowledge sharing and strengthening MITRE’s knowledge base. She can be reached at mzaharee<at>mitre.org.

What do IBM Watson, the National Oceanic and Atmospheric Administration (NOAA) National Weather Service (NWS) website and the Department of Defense (DoD) have in common? The answer is… a controlled vocabulary (CV)! Watson has roughly 200 million pages of natural language content (equivalent to 1 million books), a library of facts and a hierarchy of terms [1]. NOAA’s NWS glossary contains information on more than 2,000 terms, phrases and abbreviations used by NWS forecasters to communicate with each other [2]. The DoD has a 553-page Dictionary of Military and Associated Terms [3] that presents approved terminology for use by all DoD components. Watson, NOAA NWS and the DoD benefit from CVs because these tools

- provide a standardized list of terms used to tag data and information;
- provide a shared, commonly understood language to enable communication and knowledge exchange among stakeholders;
- build a common community of interest (COI)-wide terminology resource; and
- define content and other document types, the fields that will describe attributes and the actual values (metadata) [4].

Definitions
The term CV is often used interchangeably with taxonomy and thesaurus, but technically they are different. CVs, taxonomies and thesauri represent increasing levels of complexity, as shown in Figure 1. 

- CV – A specific list of terms for a specialized purpose. It is controlled because only terms from the list may be used for the subject area covered [6]
- Taxonomy – A collection of controlled vocabulary terms organized into a hierarchical structure with parent-child relationships [7]
Thesaurus – A controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators [7]. Several standards govern the development of CVs. The most common is American National Standards Institute (ANSI)/National Information Standards Organization (NISO) Z39-109.2005 – Guidelines for the Construction, Format and Management of Monolingual CVs [6]. This standard focuses on CVs used to represent content objects in knowledge organization systems such as lists, synonym rings, taxonomies and thesauri.

Developing a CV or Taxonomy

Figure 2 outlines a suggested approach for creating a CV or taxonomy:

- **Determine Scope.** Taxonomists assert that there is no one right way to create a CV and that taxonomies are always evolving. This assertion is true. However, it is important to meet with stakeholders to determine the project scope up front. This step may lead to production of a business case document or scope statement [8] that captures the activities covered by a project and, equally important, defines the elements that are out of scope.

- **Identify Sources.** It is essential to review a representative collection of content and to have access to subject matter experts (SMEs) in the relevant domain and to use literature from the COI domain to identify common terminology. Sources include books, journals, department terminology, databases, web resources and search logs (the latter are sometimes referred to as behavior-based taxonomies) [5]. These sources will help determine core areas for the CV/taxonomy. A key decision at this point is whether to build or buy. Can you use or modify an existing taxonomy? It is a good practice to survey existing taxonomy resources for your domain and then compare against terms gathered to-date.
**Plan for Maintenance.** A CV/taxonomy is never finished, unless it is no longer used for indexing or its database is no longer being updated. Plan for maintenance as part of development. A poorly maintained CV/taxonomy quickly becomes a liability rather than an asset [9]. Maintaining a CV/taxonomy includes changing terms, changing status of terms, deleting terms or relationships, adding new terms/relationships and even changing the structure of the CV/taxonomy [10].

**Gather Terms.** There are two main methods for building a CV/taxonomy: top down or bottom up. A top-down approach might be to convene a group of SMEs to decide on the scope and categories of terms to be included. A bottom-up method would be to assemble a set of representative, already indexed documents and use the indexed terms as a preliminary term list [9]. You can maintain list of terms in a flat or relational database.

**Categorize Terms.** Next, you should organize terms into major categories such as parent/child relationships to include related terms, preferred terms and non-preferred terms. ANSI/NISO Z39-19.2005 is an excellent resource for learning and understanding common terminology associated with organizing CVs and taxonomies.

**Manage Terms.** It is beneficial to use a commercial off-the-shelf (COTS) taxonomy management tool or database to manage terms. While there is no authoritative list of taxonomy software, many taxonomy management tools can maintain terms, their associated relationships and other attributes [6]. They include single-user desktop software, large-scale thesaurus systems, free and open source software and other software with taxonomy management components [6]. The nature of the taxonomy project determines what software product is needed.

**Visualize Terms.** A graphical representation of the taxonomy facilitates SME review and validation of subject categorization. Mind-mapping software can graphically display the categorization of a CV/taxonomy. Ontology tools can also provide a visualization capability; however, these tools require expert knowledge of ontology [11].

**Export Terms.** Terms should be extracted into a machine-readable language such as Web Ontology Language (OWL) or Simple Knowledge Organization System (SKOS).
- OWL is designed for use by applications that must process the content of information rather than merely present information to humans. The language facilitates greater machine interpretability of web content than eXtensible Markup Language (XML), Resource Description Framework (RDF) and RDF Schema (RDF-S) by providing additional vocabulary as well as a formal semantics. OWL has three increasingly expressive sublanguages: OWL Lite, OWL DL and OWL Full.
- SKOS provides a standard way to represent knowledge organization systems using the RDF [12].

**Review/Validate.** SMEs can offer invaluable assistance in reviewing terms. Develop a SME review process for reviewing both the terms and the categorization of terms. The RDF/XML from exported files should also be validated. The World Wide Web Consortium (W3C) serves as a reference source.

**Post to a Registry or Data Warehouse.** Finally, you should find a registry or data warehouse where you can post content for knowledge sharing and reuse. Your colleagues will be grateful.

**Use Case**

The DoD Net-Centric Data Strategy (NCDS) – the DoD guide to enabling data sharing in a net-centric environment – seeks to promote shared understanding within and among DoD programs. According to the strategy, the key to understandability is to create and maintain a CV.

The MITRE Corporation’s Tactical Intelligence, Surveillance, and Reconnaissance (ISR) Integration Metadata Harmonization (MDH) effort serves as an example of a project that directly supports the NCDS vision by developing a method for making data accessible and understandable to users within the intelligence community and the broader DoD. As a part of this project, MDH SMEs identify key data assets – entities composed of data, to
include databases, documents and web pages for harmonization – and map
the elements of these assets to elements in the DoD Discovery Metadata
Specification (DDMS). This process is referred to as performing a crosswalk. The SMEs then prepare a crosswalk package documenting results of the
harmonization process and post it on the DoD Metadata Registry for reuse.
Ontological products, such as a CV, taxonomy and thesaurus, complement
each crosswalk to ensure that the package meets the NCDS goal of ensuring
understandability.

Summary
As illustrated by Watson, NOAA NWS and NCDS, CVs/taxonomies are
key to making data assets widely and unambiguously understandable. Users of
CVs and taxonomies benefit by having common terms for metadata, enhancing
communication between COIs and increasing discovery of information.

Developing CVs/taxonomies is time consuming; thus, a repeatable
process is a key factor in efficiency and success. The approach proposed in
this article can serve as a guide to building useful CVs and/or taxonomies.

Resources Mentioned in the Article

MITRE is a not-for-profit corporation, chartered to work solely in the public interest, that operates multiple federally funded research and development centers (FFRDCs). An FFRDC is a unique organization that assists the U.S. government with scientific research and analysis, development and acquisition, and/or systems engineering and integration. FFRDCs address long-term problems of considerable complexity, analyze technical questions with a high degree of objectivity and provide creative and cost-effective solutions to government problems.

Approved for Public Release: 12-3989. Distribution Unlimited
©2012-The MITRE Corporation. All rights reserved
On What Assumptions Do You Base Your Thinking?
Contextual Challenges and Strategic Possibilities

by Thom Haller, associate editor for information architecture

In the mid-90s I spent time with two professional organizations, the Society for Technical Communication (STC) and a group now known as the International Society for Performance and Instruction (ISPI). STC provided me with the skills to develop usable communication products. ISPI encouraged me to think about individual and organizational performance. They reminded me that organizations often went looking for solutions (“We need training”) when they might do better by taking advantage of an alternative (“We now benefit from better instructions”).

At about the same time, I was becoming quite interested in the visual display of information and cognitive thinking. I thought I might stop calling myself a technical writer and begin calling myself a data stylist. Instead, I ran into a colleague at the 1995 STC conference who suggested that I call myself an information architect. He directed me to the work of Richard Saul Wurman, who proposed strategies for making the complex clear.

I embraced the new label and the opportunities it presented.

Meanwhile, the Internet reshaped our lives, our jobs and our communication products. “How wonderful,” I thought. “Now that we have more ways to present information, stakeholders will place more value on reducing complexity.”

Maybe not. “Let’s just get something up there,” business leaders cried. They wanted “home pages.” They wanted “features,” often based on the technology of the time. And they wanted “portals.”

Look. It’s a Portal!

During the first decade of online engagement, World Wide Web structure was typically scattered and siloed. Content emerged haphazardly; it was often program-specific, seldom reader-focused.

My teaching placed me inside an organization where mid-level managers heard my call about possibilities inherent in user-focused structure. They embraced the opportunity to improve the organization’s public-facing website, and together we reviewed content, interviewed specialists and identified core topics that were central to the organization’s mission. The mid-level employees felt they were accomplishing great success until senior stakeholders announced they would spend no money on the website and instead would invest in a portal. They envisioned an entirely new online environment where members could congregate, connect and advance the ideas of the organization. They attended to what technology
could offer them and purchased “the most complete set of capabilities and a best-of-breed solution while delivering a scalable, unified platform to handle traffic that is expected to grow exponentially in the months and years ahead.”

It didn’t work. Their public-facing web presence floundered for years. And the URL for their portal is now available for sale from GoDaddy.

**What Happened?**

Did these executives make a strategic business decision? History suggests they did not. Instead of collecting research that supported strategic decision-making, the executive team selected a product based on an imagined information structure. The tool didn’t support the tasks its visitors needed to accomplish. Nor did it provide clear pathways to the information they wanted. Ultimately, the tool failed.

The failure was, in part, contextual. The executives placed tremendous faith in a systematic solution without fully understanding the problem. Mid-level managers who better understood the problem didn’t articulate their belief in user-focused structure in such a way that mattered to decision-makers.

No strategic conversation took place.

**Our Role in Strategic Decision-Making**

As professional communicators, we are called on to develop strategies to help humans find their way through information structure. Whether or not we work on the user interface, we all work with others in mind. We know we need to fully understand our users’ perspective and translate it to and for others.

How do we translate this knowledge to stakeholders? The questions we ask are valuable to the entire organization. Consider the following:

We ask strategic questions about **audience:**
- Who are we addressing?
- What do they need to know and do?
- What do they know now?
- How do we want them to respond?

We ask strategic questions about **purpose:**
- What do our audiences want to do with our content?
- What value do they get from our content?
- What’s in it for them? Why should we matter to them?

**What is YOUR Context?**

1. Is your business motivated by the desire to “just get something up there” as opposed to the challenge of enabling people who read content to get their jobs done?
2. Are your bosses and organization motivated by preference (“Let’s develop a cool site”) as opposed to creating an online environment that supports audiences and meets an organizational mission?
3. Does your boss/organization expect a “perfect” communication product that solves a host of problems? Or do you work toward progressive success in meeting organization goals”?
4. Does your organization present itself as an expert (“the company that wrote the book on repair parts”) or as a client-focused servant (provider of communication products that enable people to make many choices themselves)?
5. Do your own perceptions of document design and construction come from “intuitive models” (as writers, we are gifted with the ability to present information in clever ways that will get the attention of our peers) as opposed to coming from a systematic understanding of our audience and a reliance on feedback models to better understand users?
6. Do your content and structure choices reflect an organization that is unsure about how it wants to present itself to its audiences or as one that provides focused and purposeful content?
We ask strategic questions about **context**:
- How do our readers respond to our mission and “value proposition”?
- What assumptions guide us?
- How do our audiences feel about our assumptions?
- What barriers do our audiences face that could preclude them from understanding or responding to what we offer?

We ask strategic questions about **priorities** and **actions**:
- What results do we seek?
- What are the priorities?
- What actions must we take?

- What are timelines and responsibilities?

Organizations have been managing online communication now for 15 years. But the culture and context of business continues to thwart what’s possible for our audiences. Ask yourself – Is your business motivated by the desire to “just get something up there?” Or is it interested in enabling people to get their jobs done?

Put on your strategist hat. Help your organization think clearly about audience, purpose, context and priorities. Help your colleagues understand how content is a critical asset. Enable them to see why good structure makes strategic sense.
We live in the age of data. As the popular statement says, “data is the new oil.” It is a growing resource with high potential. In order to realize this potential, we need to determine how to curate data effectively.

Data curation at its simplest is taking care of resources. It involves the selection, appraisal, storage and dissemination of objects and collections. With regard to data, particularly digital data, it is important to conceptualize and support the whole data lifecycle because without a proper understanding of how data are created and used it is almost impossible to store and preserve it properly. And vice versa – without adequate storage and preservation, data can quickly become obsolete and unusable. Data curation can easily become the responsibility of both everyone and no one. By understanding the data lifecycle, a data curator can alleviate this problem.

The data curator occupies a hybrid role somewhere among the traditional researcher, librarian, early technology adopter and policy maker. Clearly, neither current library and information science curricula nor extra-academic certification programs are ready to prepare people for this complex function. In an attempt to fill this gap, as well as to raise awareness of the significance of the work of such hybrid professionals, the Council on Library and Information Resources (CLIR) has expanded its postdoctoral fellowship in academic libraries to create the Data Curation Fellowship Program (www.clir.org/fellowships/postdoctoral-fellowship-in-academic-libraries-new/program-information/clir-dlf-data-curation-sciences).

The program, which is co-sponsored by the CLIR Digital Library Federation (CLIR/DLF), the Alfred P. Sloan Foundation and participating universities, is aimed at establishing data curation as a profession and providing opportunities for on-the-job training. Recent Ph.D.s already have domain-specific knowledge, as well as expertise in data collection and analysis. During this two-year postdoctoral fellowship, participants are expected to gain experience in data storage, organization, preservation and dissemination. For the 2012-2013 fellowship, there are seven fellows hosted in six institutions: B. Dewayne Branch at Purdue University; Vessela Ensberg at University of California, Los Angeles; Inna Kouper at Indiana University; Ting Wang at Lehigh University; Wei Yang at McMaster University; and Fe Consolacion Sferdean and Natsuko Hayashi Nicholls at University of Michigan.
As a fellow of this program, I participated in a two-week immersion seminar at Bryn Mawr College with other CLIR fellows (from both the data curation and academic libraries programs). The seminar brought library practitioners and researchers together to share our experiences with data and data initiatives and discuss issues facing libraries and academic institutions. The participants came from an astonishingly wide range of domains – from educational research and political sciences to geology and molecular biology. We knew that developing a common language and culture across diverse disciplines would be important, but were not clear where to begin. As someone with experience in libraries and information studies and interest in cross-disciplinary research, I was amazed by some deep disconnects between research areas, researchers and libraries, and libraries and technologies. There is a lot to be done if we want to establish and promote a common culture of data sharing.

In addition to recognizing diversity, participants discussed the changing role of the library and the challenges of developing infrastructure to support data curation. Infrastructure development is not merely a technical/engineering issue – data curators need to be aware of historical context and socio-political issues, as well as the complexities of everyday practice. Because data curation involves many different stakeholders, it is important to ease tension and facilitate consensus among universities, libraries, granting agencies and creative individuals. We also talked about collaborations and interdisciplinarity. Despite the great diversity among fellows and the uncertainties involved in data curation, everyone expressed excitement to work with different kinds of people and materials and do what is called boundary work, that is, connecting people across data types, technologies, disciplines and institutions.

It has been almost two months since the seminar ended and the fellows dispersed to their new places of work. We plan to meet online every month to reconnect, share successes and frustrations, explore potential collaborations and identify more training opportunities. From our first meeting in September, it seems that fellows have been able to engage with the challenges at work and take on the CLIR mission to raise awareness and build capacities for data management and sharing throughout the academy. As CLIR program officer Christa Williford kindly noted, “This cohort of data curation fellows has exceeded our expectations with their level of engagement, problem-solving skills and openness to new ideas. We are learning a great deal from them.”

Along with this promising beginning come many uncertainties. We chose an unusual, trailblazing path in our academic careers. We do not know whether this path will allow us to acquire enough symbolic capital to be recognized as serious researchers. We also do not know whether data curation as a profession will gain high, Ph.D.-required status or whether some of us will have to choose between traditional research and professional careers. We also face the challenges that all academics face – balancing research, teaching and service activities; making an impact; and living a productive life. In one of next year’s RDAP columns, we will address some of those issues and provide a more detailed report about our experiences as hybrid data curation scholars.